DOCUMENT RESUME

ED 073 121 TM 002 363

AUTHOR	Crooks, Lois A., Ed.
TITLE	An Investigation of Sources of Bias in the Prediction of Job Performance. A Six-Year Study.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Civil Service Commission, Washington, D.C.; Ford Foundation, New York, N.Y.
PUB DATE	22 Jun 72
NOTE	118p.; Proceedings of an Invitational Conference held June 22, 1972, New York, New York
AVAILABLE FROM	Educational Testing Service, Rosedale Road, Princeton, New Jersey 08540 ($3.00)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Aptitude Tests; Conference Reports; Correlation; Criterion Referenced Tests; Data Analysis; Data Collection; Ethnic Groups; Government Employees; Minority Groups; Multiple Regression Analysis; *Performance Criteria; Personnel Selection; Predictive Validity; Questionnaires; Rating Scales; Research Methodology; Speeches; Tables (Data); *Task Performance; *Test Bias; Test Results; *Test Validity

ABSTRACT

This invitational conference was convened to report the principal findings of a six-year study of possible sources of bias in the prediction of job performance. The research was conducted jointly by Educational Testing Service and the U.S. Civil Service Commission, supported by the Ford Foundation. Data were gathered on test and job performance of ethnic subgroups in three occupations in the Federal Government. The design of the study permitted a detailed analysis of the differential validity of selected aptitude tests for several kinds of performance criteria. Speakers at the conference were asked to respond to a draft of the technical report, to be published in 1973. Following an introduction to the project and a presentation of the major findings, the papers are provided. The titles and authors of the papers are as follows: "Technical Critique" by Anne Anastasi, "Implications for Employers in Government" by Raymond Jacobson, "Implications for Employers in Industry" by Lewis E. Albright, "Implications for Blacks" by Roscoe C. Brown, Jr., "Implications for Spanish Americans" by Edward J. Casavantes, "Implications for Governmental Regulatory Agencies" by Robert M. Guion, and "Implications for Future Research" by S. Rains Wallace. (Author/DB)

FILMED FROM BEST AVAILABLE COPY

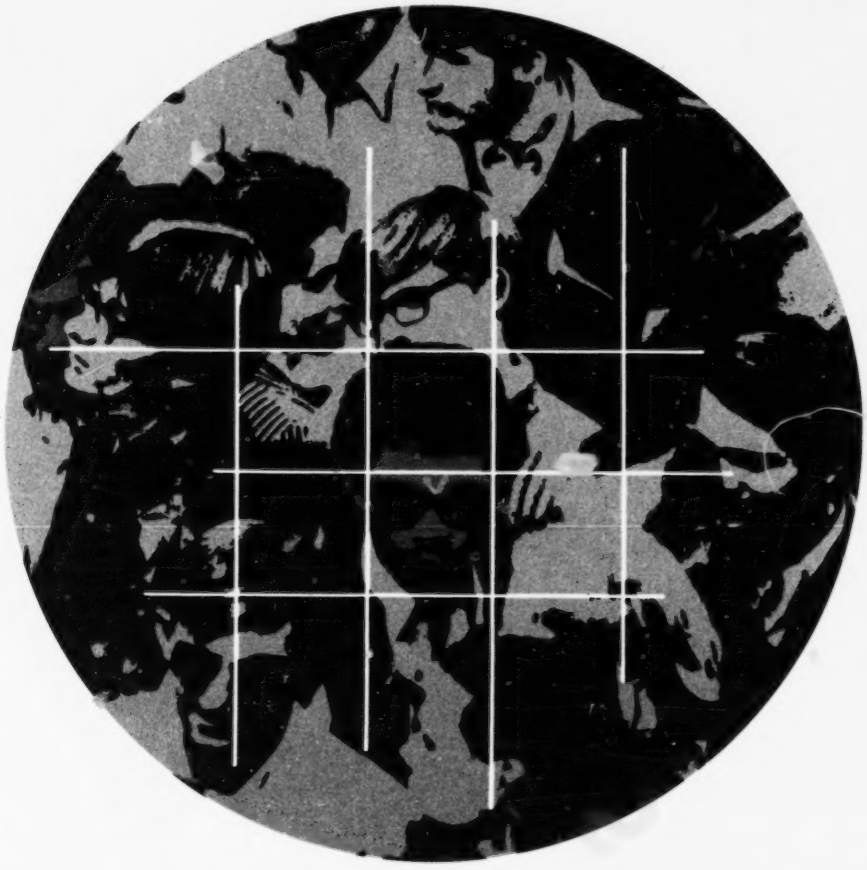
U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 073121

TM 002 063

An Investigation of Sources of Bias in the Prediction of Job Performance

A Six-Year Study



PROCEEDINGS OF INVITATIONAL CONFERENCE
The Barclay Hotel
New York, New York
June 22, 1972

EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY

ED 073121

AN INVESTIGATION OF SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE

. . . A SIX-YEAR STUDY . . .

Proceedings of an Invitational Conference

held June 22, 1972

The Barclay Hotel

New York, New York

Lois A. Crooks, Editor

Additional copies of this report may be obtained

from: Educational Testing Service
Rosedale Road
Princeton, New Jersey 08540

Price per copy: $3.00

PERMISSION TO REPRODUCE THIS COPY RIGHTED MATERIAL HAS BEEN GRANTED BY
ETS
Lois Cooks 1/9/73
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

Copyright © 1972 Educational Testing Service, Princeton, N. J.

FOREWORD

This invitational conference was convened to report the principal findings of a six-year study of possible sources of bias in the prediction of job performance. The research was conducted jointly by Educational Testing Service and the U. S. Civil Service Commission, supported by the Ford Foundation.

Data were gathered on test and job performance of ethnic subgroups in three occupations in the Federal Government. The design of the study permitted a detailed analysis of the differential validity of selected aptitude tests for several kinds of performance criteria.

Because the findings have implications for employers, behavioral scientists, and others concerned with social and public policy issues, invited speakers in these areas were asked to respond to a draft of the technical report, which will be published early in 1973. Their papers are included here, following an introduction to the project and a presentation of the major findings.

Many people have been involved over the past six years in the design and direction of the project, in the development of the instrumentation, and in data collection and analyses. The study could not have been carried out without the assistance of those in the Federal agencies who facilitated the data collection, and the cooperation of the 1,400 job incumbents who were the subjects.

Members of the Advisory Committee, who were convened periodically for consultation on research design, progress of the study, and implications of the findings, filled an invaluable role. They were:

John K. Hemphill, Far West Laboratory for Educational Research and Development, Chairman

Marvin D. Dunnette, University of Minnesota

Robert M. Guion, Bowling Green State University

S. O. Roberts, Fisk University

Members of the Management Committee, who joined with the Advisory Committee

in following the progress of the study, made themselves available for counsel and support on a day-to-day basis. Their names, with period served, follow:

William W. Turnbull, President, Educational Testing Service (until July, 1969)

Albert P. Maslow, Chief, Personnel Measurement Research and Development Center, U. S. Civil Service Commission (until September, 1971, when he joined Educational Testing Service)

Samuel J. Messick, Vice President, Educational Testing Service (from July, 1969)

William A. Gorham, Associate Director, Personnel Measurement Research and Development Center, U. S. Civil Service Commission (from September, 1971)

The Project Staff included William A. Gorham (then Chief of Research and Development) until he joined the Management Committee, and Mary L. Tenopyr, now Chief of Research, from the U. S. Civil Service Commission, and the following from Educational Testing Service. Those marked with asterisks worked on the study from its inception to completion. The other staff members were involved at various stages of the study, as it progressed.

*Joel T. Campbell, Senior Research Psychologist, Principal Investigator
*Donald A. Rock, Senior Research Psychologist
Franklin R. Evans, Research Psychologist
Ronald L. Flaugher, Research Psychologist
Lewis W. Pike, Research Psychologist
David M. Nolan, Director, Washington, D. C. Office
Lois A. Crooks, Associate Research Psychologist
William S. Hall, Associate Research Psychologist
Lila Norris, Associate Research Psychologist
Barbara Dynarski, Senior Research Assistant
*Margaret H. Mahoney, Senior Research Assistant
Mary Ellen Parry, Senior Research Assistant
Harriet Blizzard, Research Assistant
*Virginia Rau, Administrative Assistant
C. Brooke Scaramozzino, Secretary

Other ETS staff members, not listed, assisted in data collection.

William A. Gorham

Samuel J. Messick
Conference Co-Chairmen

Invitational Conference on
Sources of Bias in the
Prediction of Job Performance

Barclay Hotel, 111 East 48th Street
New York City
June 22, 1972

Program

William A. Gorham, U. S. Civil Service Commission — Co-Chairmen — Samuel Messick, Educational Testing Service

9:00	Registration and coffee	
9:30	Background and Design of the Project	Albert P. Maslow Educational Testing Service
9:50	Principal Project Results and Conclusions	Joel T. Campbell Educational Testing Service
10:45	Recess	
11:00	Technical Critique	Anne Anastasi Fordham University
	Implications for Employers in Government	Raymond Jacobson U. S. Civil Service Commission
	Implications for Employers in Industry	Lewis E. Albright Kaiser Aluminum and Chemical Company
12:00	Lunch	
1:30	Implications for Blacks	Roscoe C. Brown New York University
	Implications for Spanish-Americans	Edward J. Casavantes Association of Psychologists for La Raza
2:15	Recess	
2:30	Implications for Governmental Regulatory Agencies	Robert M. Guion Bowling Green State University
	Implications for Future Research	S. Rains Wallace The Ohio State University
3:15	General Discussion	
4:00	Cocktails	

CONFERENCE PARTICIPANTS

Lewis E. Albright
Kaiser Aluminum and
Chemical Corporation

Anne Anastasi
Fordham University

Richard S. Barrett
Hastings-on-Hudson, New York

Claude J. Bartlett
University of Maryland

V. Jon Bentz
Sears, Roebuck and Company

Virginia R. Boehm
New York State Division
of Employment

Douglas W. Bray
American Telephone and
Telegraph Company

Roscoe C. Brown, Jr.
New York University

Joel T. Campbell
Educational Testing Service

Edward J. Casavantes
Association of Psychologists
for La Raza

Lois A. Crooks
Educational Testing Service

Frederick B. Davis
University of Pennsylvania

Robert D. Dugan
International Telephone and
Telegraph Company

Marvin D. Dunnette
University of Minnesota

Patricia Ann Dyer
IBM Corporation

William H. Enneis
U. S. Equal Employment
Opportunity Commission

Ronald L. Flaugher
Educational Testing Service

Edmund F. Fuchs
U. S. Army Behavioral Sciences
Research Laboratory

William A. Gorham
U. S. Civil Service Commission

Donald L. Grant
American Telephone and
Telegraph Company

Robert M. Guion
Bowling Green State University

John Howland
U. S. Civil Service Commission

Raymond Jacobson
U. S. Civil Service Commission

Clifford E. Jurgensen
Minneapolis Gas Company

Charles H. Lawshe
Purdue University

Samuel Leff
U. S. Civil Service Commission

Roger T. Lennon
Harcourt Brace Jovanovich

Robert L. Linn
Educational Testing Service

Margaret H. Mahoney
Educational Testing Service

Albert P. Maslow
Educational Testing Service

Samuel J. Messick
Educational Testing Service

CONFERENCE PARTICIPANTS (continued)

Herbert H. Meyer
General Electric Company

Jean Palermo
Science Research Associates

Lawrence Plotkin
The MARC Corporation

Ersa Poston
New York State Department
of Civil Service

S. O. Roberts
Fisk University

Donald A. Rock
Educational Testing Service

Morton Rosen
New York State Department
of Labor

Floyd L. Ruch
Psychological Services, Inc.

Robert J. Solomon
Educational Testing Service

C. Paul Sparks
Humble Oil & Refining Company

Erwin K. Taylor
Personnel Research &
Development Corporation

Mary L. Tenopyr
U. S. Civil Service Commission

William W. Turnbull
Educational Testing Service

S. Rains Wallace
The Ohio State University

E. Belvin Williams
Educational Testing Service

TABLE OF CONTENTS

	Page
Foreword	i
Agenda for the Conference	iii
Roster of Participants	v
Table of Contents	vii
BACKGROUND AND DESIGN OF THE PROJECT Albert P. Maslow	1
PRINCIPAL RESULTS OF THE STUDY AND CONCLUSIONS Joel T. Campbell	9
TECHNICAL CRITIQUE Anne Anastasi	79
IMPLICATIONS FOR EMPLOYERS IN GOVERNMENT Raymond Jacobson	89
IMPLICATIONS FOR EMPLOYERS IN INDUSTRY Lewis E. Albright	95
IMPLICATIONS FOR BLACKS Roscoe C. Brown, Jr.	101
IMPLICATIONS FOR SPANISH-AMERICANS Edward J. Casavantes	111
IMPLICATIONS FOR GOVERNMENTAL REGULATORY AGENCIES Robert M. Guion	129
IMPLICATIONS FOR FUTURE RESEARCH S. Rains Wallace	137

BACKGROUND AND DESIGN OF THE PROJECT

Albert P. Maslow

Director, Government & Professional Programs

Educational Testing Service[1]

This project grew out of a series of meetings between staff of the Educational Testing Service and the Civil Service Commission, under the leadership of Mr. Chauncey and Mr. Macy, in which the two organizations were exploring areas of mutual interest and possible cooperation. At that time, 1965, concern for the fairness of testing practices was a major topic on many fronts. Discontent with tests as a perceived barrier to selection and promotion of employees, both in industry and government, was widespread among minority groups.

The then-current report of the National Advisory Commission on Civil Disorders stated that existing testing procedures should either be "revalidated or replaced by work samples or job tryouts." At that time, also, the Office of Federal Contract Compliance was developing test regulations attempting to assure that tests were validated and were in effect color-blind in each particular job situation.

In these staff discussions there was complete agreement, of course, that these concerns were legitimate and that the objective of improving the use of tests was desirable. There was a nagging worry, however, that the various proposals for replacing tests or modifying their use would exacerbate rather than reduce discrimination in hiring practices.

Some felt that because of costs and technical hazards, empirical validation of tests would be found infeasible. Such a conclusion would result in

[1] Prior to September, 1971, Dr. Maslow was Chief, Personnel Measurement Research and Development Center, U. S. Civil Service Commission.

employers abandoning tests in favor of other selection practices, such as interviews, which would in effect be less objective and more open to bias, intended or unintended. Too, there was the danger that if employers were to abandon tests of an aptitude nature and seek refuge in prescriptions of experience and training for particular jobs, the effect would be to lock out minorities even more strongly. It is just that group which has been least able to gain job experience. The available research literature on these issues at that time was scanty.

In designing a research plan to propose to the Ford Foundation, we confronted the question of whether it was really practicable to conduct research of the scope that would have a chance of throwing some clear light on these problems. Could, in fact, sizeable groups of majority and minority employees be located such that they were in similar jobs and under common supervision, had followed similar career paths in reaching their current jobs, and had not been directly screened by employment tests, so that restriction in range would not be fatal to the research? Could we expect to find, or develop, a variety of measures of job performance so that the validity of tests for different criteria could be investigated? Finally, there was concern as to whether we could expect the cooperation of agency management and supervisors and of the employees themselves for the heavy commitment of time and interest demanded.

After a check of occupational data and other considerations we found that the conditions for a sound study seemed to exist in the Veterans Administration in the occupation of Medical Technician. Here it was possible to locate and eventually study 168 Black and 297 Caucasian employees in some 30 hospitals across the country.

The research design was straightforward. Intensive job analysis by a variety of techniques was made in a wide sampling of installations. From these

job analyses, a careful selection of tests was made to measure the aptitude and ability factors considered critical to job performance. The source of these tests was mainly the French, et al., kit of factored tests.[1] One Civil Service test was used. (In later studies, other Civil Service tests and one from the Flanagan Industrial series were also used in addition to selected tests from the French kit.)

A detailed questionnaire was designed to develop information on the personal history of all of the study groups. This questionnaire covered the obvious biographical data, education and experience data, and training on and off the job. In the area of work performed, a detailed task checklist was developed from interviews and observations of the job to determine the intensity, importance, and relative complexity of the tasks performed by the job incumbents.

Performance is multi-dimensional. Accordingly, three (3) types of performance measures were developed. These included specially constructed rating scales, defined and anchored by behavioral descriptions of aspects of job performance, a work sample, and a job knowledge test.

As a part of the feasibility study, a special effort was made to inform and to solicit cooperation and understanding of groups outside of the research staffs. Meetings were held with employee union representatives, with representatives of key minority groups, and with management personnel at the Veterans Administration to discuss the purposes and objectives of the study and to invite their cooperation as appropriate. The usual precautions were taken to pretest all of the instruments before final use. Special precautions also were taken to advise the employees who were asked to take part in the

[1] French, J. W., Ekstrom, R. B., & Price, L.A. Kit of reference tests for cognitive factors. Princeton, New Jersey: Educational Testing Service, 1963.

study of its purpose and of the confidentiality of the data, and a plan was set up to report back to them information concerning their test performance.

The study with Medical Technicians was then conducted according to this plan and in March of 1968, when the Project Advisory Committee met to consider the results of this first effort, the Committee came to the unanimous conclusion that:

> The study as conducted to date has demonstrated conclusively the feasibility of proceeding with the major investigation described in the original proposal. It has also served to demonstrate the enormous technical and logistical difficulty of conducting the work, and to suggest new approaches that should be incorporated in the research. We are more deeply than ever convinced, however, that the study proper is not only feasible but of major importance for social policy in the employment of minority and majority group members, both in government service and in the private sector. If carried out fully at the level of thoroughness and competence demonstrated in the work to date, the investigation promises to stand as a landmark study.
>
> The implications of the study may well be critical for the effective and equitable operation of employment systems based on the recognition of merit. It is therefore of first importance that the full investigation provide for a level of effort commensurate with its potential significance.

The research staff issued a series of technical reports on the Medical Technicians study. The first general report was made in a symposium at the annual convention of the American Psychological Association in September, 1969. At that time, Dr. Campbell and others of his staff presented technical reports on the findings of the feasibility study. We were encouraged by the comments of the discussant at that symposium, Dr. Mary Tenopyr, who felt that the design, the analyses, and the tentative interpretations of the feasibility study were very sound and provided a model appropriate to further research in this program.

For the second phase of the study, the occupation of Cartographic Technician was used. This was particularly suitable because it included not only Caucasian and Black employees, but also a large group of Mexican-Americans.

These employees were found in the U. S. Department of the Army (Corps of Engineers and Topographic Command) and the U. S. Department of Commerce (Coast & Geodetic Survey and Census Bureau). This occupation also provided a chance to see whether the findings for technicians in the medical field would replicate for technicians in Cartography. The same general plan was followed as for the Medical Technicians.

The final study was made with the occupation of Inventory Management Specialist. These employees are primarily in the U. S. Department of Defense agencies, and include a substantial number of Blacks and Mexican-Americans. This occupation was somewhat different from the two technician occupations in that employees could enter it from a variety of lower level technical and clerical jobs as well as directly from outside the service, and in that it had a longer career ladder reaching into middle and top management and professional positions. Thus, it appeared to require a somewhat different set of skills and abilities that would broaden the scope of the research. For this particular occupation, it was possible to develop a work sample as for the technician jobs, but it did not appear to be feasible to develop a job knowledge test because of the varying nature of the procedures and the materiel managed across agencies and installations. Therefore, evaluation of job knowledge was made from data obtained through supervisory ratings and work sample procedures.

The analysis of the data followed a common pattern in each study. Briefly, the steps were:

1) To examine the background data and task analyses to see whether any systematic differences exist among the ethnic groups.
2) To examine and compare the performance of ethnic groups on each of the predictor measures.
3) To examine ethnic group differences on job performance measures.

This involved several steps:

a) A study of the interrelation of the performance measures to see whether they reflect different aspects of job performance.

b) A study of whether the performance measures might have different values for different ethnic groups.

4) A major issue, of course, is whether tests are differentially valid for different ethnic groups and, more to the point of job bias, whether regression lines differ significantly for ethnic groups. If this were so, the same test score would lead to different predictions for a job applicant depending on his ethnic group. Or, to use Guion's definition: "two people with equal test scores could then have an unequal probability of being hired."

To resolve these issues, analyses were made to compare the validities of separate measures by ethnic group, and to compare the regression lines for each of the predictor measures for the several ethnic groups.

5) Among the ideas advanced in recent years is the notion that different prediction equations for ethnic groups may be needed to counteract test bias. Such a conclusion would, of course, present serious policy and perhaps legal problems, especially for merit systems. It could lead, for example, to a different set of tests for one group than for the other. It could also lead to different scoring, weighting, and ranking procedures on the same tests for different groups.

The design, therefore, included an analysis of the

differences in the multiple regression equations for the separate ethnic groups, and a study of the effect of using the regression equation for one ethnic group to predict performance for another group.

6) Finally, as we might have expected, the study very early uncovered an unexpected problem. This grew out of observations of the different effects on supervisory ratings when different ethnic combinations of rater and ratee were studied.

The implications of this kind of outcome on the policies and practices as to the use of supervisory ratings as a major personnel tool are quite obvious. This interactive effect was made the subject of a special analysis.

This neat outline of the study should not obscure the fact that the Advisory Committee and research staff confronted many very troublesome questions, both conceptual and analytical. They experienced the "enormous complexity" of such a research effort, and the fact that the studies have been pushed to completion is, I think, the best testimony to their competence and dedication.

But in all fairness, not everything was rosy. It seems that in such a sensitive area, pretest eventually leads to protest. A number of minority employees at one installation did, in fact, refuse to cooperate, and walked out of the testing room. While in this one case it did not cripple the research design, the incident does raise some disturbing questions for future research of this kind.

PRINCIPAL RESULTS OF THE STUDY AND CONCLUSIONS

Joel T. Campbell

Senior Research Psychologist

Educational Testing Service

From the description Dr. Maslow has given you of the data gathered for the project, you can well imagine that the data analysis has been extensive. Correlation coefficients and standard deviations have poured forth by the bucketful!

This morning I shall try to give the "essence" from these analyses. We will be looking at several different aspects of the data.

First, we will compare, across ethnic groups, some of the personal background and experience variables.

Next we will compare mean performance on aptitude tests and on criterion measures.

Correlation of aptitude tests with different kinds of criteria will be our next consideration, followed by comparisons of regression lines.

We will then look at multiple correlation and cross-ethnic cross-validation. Finally, we will consider the effect on ratings of ethnic group rater-ratee interaction.

Table 1[1] shows some of the background variables for the Medical Technicians. We had thought beforehand that we might find that members of one ethnic group had much shorter job experience than the other, or perhaps much less education. As you can see in Table 1, that is not what we found. There are some differences, but these are less than expected.

Table 2, for the Cartographic Technicians, gives us a similar picture. The Mexican-Americans show some differences from the other two groups, but on the whole, the background variables are very similar.

[1] Tables and figures appear at the end of this paper in the order discussed.

Table 3, for Inventory Managers, also shows very similar patterns for all three groups.

One place where we did not find similar patterns was in response to the task lists for Inventory Managers. Here, the responses for Mexican-Americans appeared to be quite discrepant. Further exploration showed that the real difference was between those working in San Antonio versus those working elsewhere, even though all were classified as Inventory Managers. Table 4 illustrates this point. (As we shall see later, this difference also affected some of the subsequent analyses.)

Comparison of mean scores on aptitude tests and criteria across ethnic groups

We next turn to a comparison of mean scores on aptitude tests and criterion measures. To do this quickly, we have plotted the minority group means as standard score departures from the Caucasian group means. All of the aptitude tests are plotted, and three of the rating scales: Learning Ability, Job Knowledge, and Overall Job Performance. Also plotted, where available, are the Job Knowledge Tests and Work Samples.

(I should mention here that the standard deviations are quite similar across ethnic groups, both for aptitude tests and criterion measures.)

Figure 1 shows the data for Medical Technicians, Figure 2 for TOPOCOM Cartographic Technicians, Figure 3 for Coast & Geodetic Survey Cartographic Technicians, and Figure 4 for Inventory Managers.

In these figures we can see that the minority groups generally score about one-half standard deviation below the Caucasian mean on aptitude tests. This difference is also reflected for the objective criterion measures (Job Knowledge Tests and Work Samples) but not for the rating scales.

Test validity

We shall next consider the very important question of the validity of the

tests for different groups against the different kinds of criteria.

Table 5 shows, for TOPOCOM Cartographic Technicians, the validity coefficients against Learning Ability ratings and Overall ratings. It can be seen that the coefficients are overwhelmingly positive and, with a few exceptions, significantly different from zero. You will also notice that the validities are usually higher using the Learning Ability rating as the criterion. These same observations apply to the other tables of correlations between tests and ratings.

As an example of correlations between tests and an objective criterion we can look at Table 6. The correlations between tests and the Work Sample score for Inventory Managers are all positive, mostly significantly different from zero, and somewhat larger than the validities against rating scales. These observations also apply to the other tables showing validities against Job Knowledge Test scores and Work Samples.

To show as much information as quickly as possible, we have plotted, for each job grouping and for each ethnic group, test validities against the Learning Ability rating, the Job Knowledge Test, and the Work Sample.

Figure 5 shows the validity coefficients for the Medical Technicians. As you can see, the patterns are very similar for both ethnic groups. In this and the succeeding figures, there usually are only a few points of difference between the validity coefficient for one ethnic group and that for another. A test which is valid for one ethnic group is usually valid for others, and conversely, a test not valid for one ethnic group lacks validity for all.

Figure 6, for TOPOCOM Cartographic Technicians, shows this pattern particularly clearly. Also noteworthy is the extent to which validity (or lack thereof) is reflected across criterion measures.

Figure 7, for the Coast & Geodetic Cartographic Technicians, shows

validities against two ratings scales. The patterns here reflect those of Figure 6 rather well.

Figure 8 shows the validity coefficients for Inventory Managers against Learning Ability rating and Work Sample score. Again, the pattern of "valid for one group - valid for all" holds pretty well here. There perhaps are more differences here from one criterion measure to another, most notably for the validities associated with the vocabulary tests.

Regression line comparisons

Our next consideration will be the comparison of regression lines.

Table 7 shows the Gulliksen-Wilks comparisons for the aptitude tests against Overall ratings, Job Knowledge Tests, and Work Samples. With ratings as the criterion, very few comparisons had significant differences. With the Job Knowledge Test as the criterion, most of the comparisons had significant differences in the intercept. In these instances, the regression line for Caucasians was above that for the minority group.

With the Work Sample as the criterion, there were again significant differences for most of the regression lines on one or another aspect of the analyses, as you can see in Table 7. For the Cartographic Technicians, and for the Black Inventory Managers, the differences--where they existed--were as before in favor of the minorities. However, for the Mexican-American Inventory Managers, the regression lines were above those for the Caucasians. In nine out of 12 instances, the difference was not statistically significant. In the other three instances with significant differences in the dispersions, it is inconclusive whether the location of the lines would be significantly different. Nevertheless, this does appear to be an instance where these three tests may be biased against one of the minority groups in predicting the Work Sample criterion.

You will recall, however, that we did find that the job patterns in the San Antonio installation appeared to be different from those in other installations. This finding raised several questions. How would the regression lines for the Mexican-Americans compare with San Antonio Caucasians? Similarly, what about the comparison of Blacks and Caucasians at the other installations? These comparisons are shown in Table 8. Now the differences between the Mexican-Americans and the Caucasians with regard to the Work Sample criterion disappear. However, we now find that the rating criterion produces differences between the Blacks and Caucasians. The difference is significant for only two out of 12 regression lines. This is another instance of apparent bias against a minority group for two tests in predicting one of the criteria.

For those of you who like a visual presentation, we have selected four figures showing regression line comparisons. Figure 9 shows a situation where there is no statistical difference between the regression lines, and no apparent difference either. Figure 10 illustrates a significant difference in slope. Figure 11 illustrates a significant difference in slope between the Caucasian and Mexican-American regression lines and a significant difference in intercepts between the Caucasian and Black lines. In Figure 12, there is no significant difference between the Mexican-American and San Antonio Caucasian regression lines. Between the Philadelphia, Dayton, and Detroit Black and Caucasian lines, there is a significant difference in intercepts, favorable to the Blacks.

The apparent difference in slopes between the lines for the two Caucasian samples is perhaps particularly noteworthy.

Another way of looking at the same kinds of relationships is shown in Table 9. In this contingency table we can see that the scores on the Map Planning Test (which best predicted Supervisors' Overall Rating for Caucasians) produces valid discrimination generally for all three ethnic groups and for all three criteria.

Similarly, Table 10--for the Subtraction & Multiplication Test for Inventory Managers--shows generally valid discriminations for the different groups. (Caucasians from different installations are lumped together here, not broken out as in Figure 12.)

Multiple correlation and regression

Our next concern will be with what happens when several predictors are combined in a multiple regression equation.

Table 11 compares predicted criterion scores for the two minority groups for multiple regression equations computed for the minority samples, and computed for the Caucasian samples.

You can see that Black subjects with high test scores are better off (receive higher predicted scores) if a regression equation derived on Black samples is used. Those with average or low test scores are better off if the regression equation derived on Caucasian samples is used.

The Mexican-Americans show a slightly different picture. Here, those with high or average scores are better off if the Mexican-American regression equation is used, while there appears to be little difference for those with low scores, whether the Mexican-American or Caucasian equations are used.

The level of accuracy of prediction can be shown by plotting multiple correlation coefficients and cross-ethnic cross-validation coefficients on the same chart. Figure 13 shows this comparison for Black and Caucasian samples from the different occupations, and Figure 14 shows similar comparisons for Mexican-American and Caucasian samples. In each figure, the distance between each point and the diagonal line represents the loss in prediction from using regression weights from a different ethnic group (that is, Caucasian regression weights for a Black sample, or vice versa). In general, very similar multiples are obtained.

Effects of rater-ratee ethnic group interaction

Our final concern is to look at what happens to ratings when a job incumbent from one ethnic group is rated by a supervisor of his own ethnic group and by a supervisor from another ethnic group.

For each of the job samples, we prepared tables like Table 12 showing the mean ratings of ethnic incumbents by ethnic supervisors.

These tables have been summarized in Table 13. The tendency of supervisors to give higher ratings, on the average, to members of one's own ethnic group comes through rather clearly here.

The correlation of Learning Ability ratings with all of the objective measures for each of the rater-ratee combinations for Cartographic Technicians is shown in Table 14. This, and similar tables for the other occupations, are summarized in Table 15.

Here we find that we have higher "validities" where Black supervisors have rated Black job incumbents than when these supervisors have rated Caucasians. Mexican-American and Caucasian raters tend to produce higher validities when rating members of other than their own ethnic group.

Table 16 shows the average correlation coefficients for the different combinations. Here, the high validity overall for the Black-rating-Black combination is striking. The correlation resulting from Mexican-American supervisors rating Caucasians is almost as high, but is based on only one sample.

Finally, Figure 15 shows the regression lines for predicting job knowledge ratings from Job Knowledge Test scores for the different combinations in the Medical Technicians study. Note the difference in the regression lines for Blacks rated by Black supervisors and Blacks rated by Caucasian supervisors. Also, the difference should be noted in the two lines for Caucasian ratees.

Obviously, in this study we have not explored all of the variables that can affect rating behavior, but I think there is little doubt that ethnic group of rater and ratee does make a difference.

Summary

A few main points should perhaps be reiterated in summary.

First, aptitude tests which have validity in relation to job performance for one ethnic group generally show validity for other ethnic groups as well.

Second, tests which are valid against a rating criterion also show validity against more objective criterion measures.

Third, multiple regression weights determined on a single ethnic group hold up surprisingly well on cross-validation across different ethnic groups.

Fourth, ethnic group rater-ratee combinations interact to affect the ratings assigned, but the effect appears to be complex and probably differs from one ethnic group to another.

Table 1 - Background Information for Medical Technicians

		Percent	
		Black	Caucasian
Sex	Male	46	47
	Female	54	53
Age	60 +	2	2
	50 - 59	8	19
	40 - 49	29	31
	30 - 39	43	22
	20 - 29	18	25
	Less than 20	0	1
Education	Advanced study	5	2
	College degree	8	7
	College, more than 2 years	21	18
	College, 2 year terminal	7	5
	College, less than 2 years	32	31
	High school graduate	20	31
	Some high school	4	4
	8th grade or less	0	1
Source of training as Medical Technician	Accredited school	40	31
	Military service	17	28
	Government hospital	23	11
	Civilian hospital	7	13
	Civilian laboratory	5	6
	Other	6	10
Total years of experience	Over 20	8	25
	16 - 19	14	12
	12 - 15	21	16
	8 - 11	21	16
	4 - 7	18	17
	2 - 3	5	6
	Less than 2	10	8
Salary grade (GS) level	8 or higher	4	5
	7	21	20
	6	36	41
	5	27	24
	4 or lower	12	10

Table 2 - Background Information for Cartographic Technicians

		Percent		
		Black	Mexican-American	Caucasian
Sex	Male	62	79	34
	Female	38	21	66
Age	60 +	2	0	2
	50 - 59	6	9	13
	40 - 49	39	27	23
	30 - 39	36	62	26
	20 - 29	18	2	36
Education	1 or more year graduate	0	0	1
	3 or 4 years college	20	1	5
	1 or 2 years college	41	27	25
	Tech or Voc institute	15	13	18
	11th or 12th grade	24	56	50
	9th or 10th grade	0	2	2
	8th grade or less	0	1	0
Total years of experience	20 or more	4	3	4
	16 - 19	20	7	15
	12 - 15	21	27	13
	8 - 11	13	34	14
	4 - 7	26	27	37
	2 - 3	14	1	14
	Less than 2	1	0	3
Salary grade (GS) level	12	1	0	0
	11	5	0	8
	10	0	0	0
	9	52	83	55
	8	10	0	8
	7	32	17	23
	6	0	0	0
	5	0	0	5

Table 3 - Background Information for Inventory Managers

		Percent		
		Black	Mexican-American	Caucasian
Sex	Male	37	62	52
	Female	58	32	38
	No response	5	5	10
Age	60 +	5	0	6
	50 - 59	25	11	33
	40 - 49	42	50	29
	30 - 39	18	31	13
	20 - 29	4	3	9
	No response	6	5	10
Education	Graduate school	3	3	2
	3 or 4 years college	20	15	22
	1 or 2 years college	28	26	9
	1 or 2 years Tech or business institute	17	9	19
	11th - 12th grade or GED diploma	27	41	34
	9th - 10th grade	0	3	4
	8th grade or less	0	1	0
	No response	5	3	10
Salary grade (GS) level	11	21	8	20
	9	67	68	66
	7	6	19	4
	No response	6	5	10
Years of experience as Inventory Manager	20 or more	4	0	4
	16 - 19	5	3	6
	12 - 15	18	15	10
	8 - 11	23	26	19
	4 - 7	28	34	38
	2 - 3	11	12	8
	Less than 2	5	5	5
	No response	7	5	10

Table 4

COMPARISONS OF SELECTED TASKS FOR INVENTORY MANAGEMENT SPECIALISTS

WITHIN AND ACROSS LOCATIONS STUDIED

	Percent reporting "significant part of job every day" or "substantial part of job, at least several times a week"			
	Black	Mexican-American (San Antonio)	Caucasian (San Antonio)	Caucasian (All Locations)
Authorize purchase of additional quantity of materiel over amount normally used	25	5	4	19
Work with Procurement in cancelling contracts	15	3	1	10
Make manual buy for urgent request	54	3	7	37
Use mathematical formulas to figure net assets	40	18	19	29
Request expediting action to assure on-time delivery of item	61	42	30	47
Arrange for transfer of stock from one depot to another	35	18	18	26
Request Cataloging Branch to assign Federal Stock Numbers to items and place in FSN Catalog	8	22	22	9
Submit back order status report	33	15	19	20
Cancel purchase orders	42	15	18	23

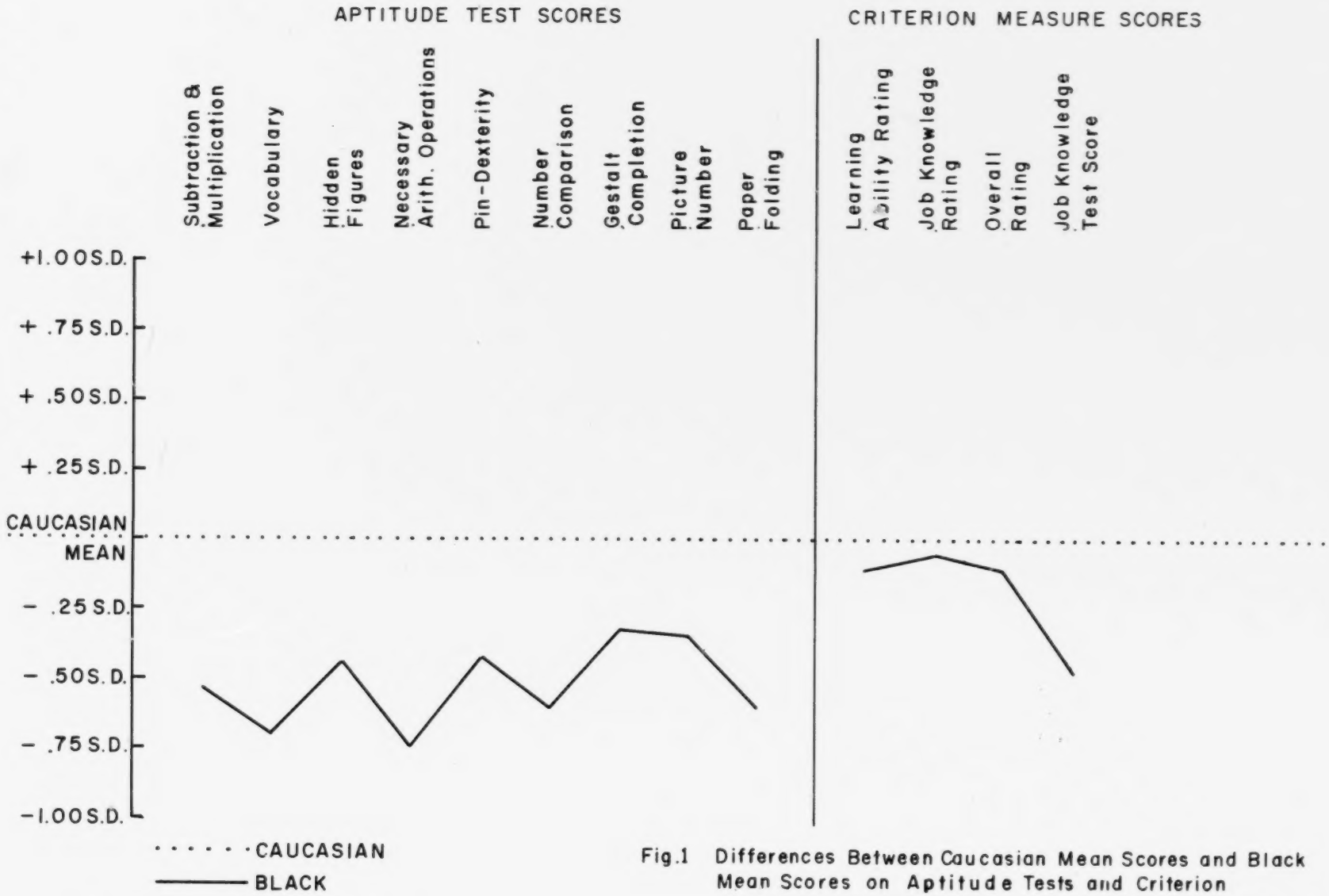


Fig.1 Differences Between Caucasian Mean Scores and Black Mean Scores on Aptitude Tests and Criterion Measures in Terms of Caucasian Standard Deviation Units MEDICAL TECHNICIANS

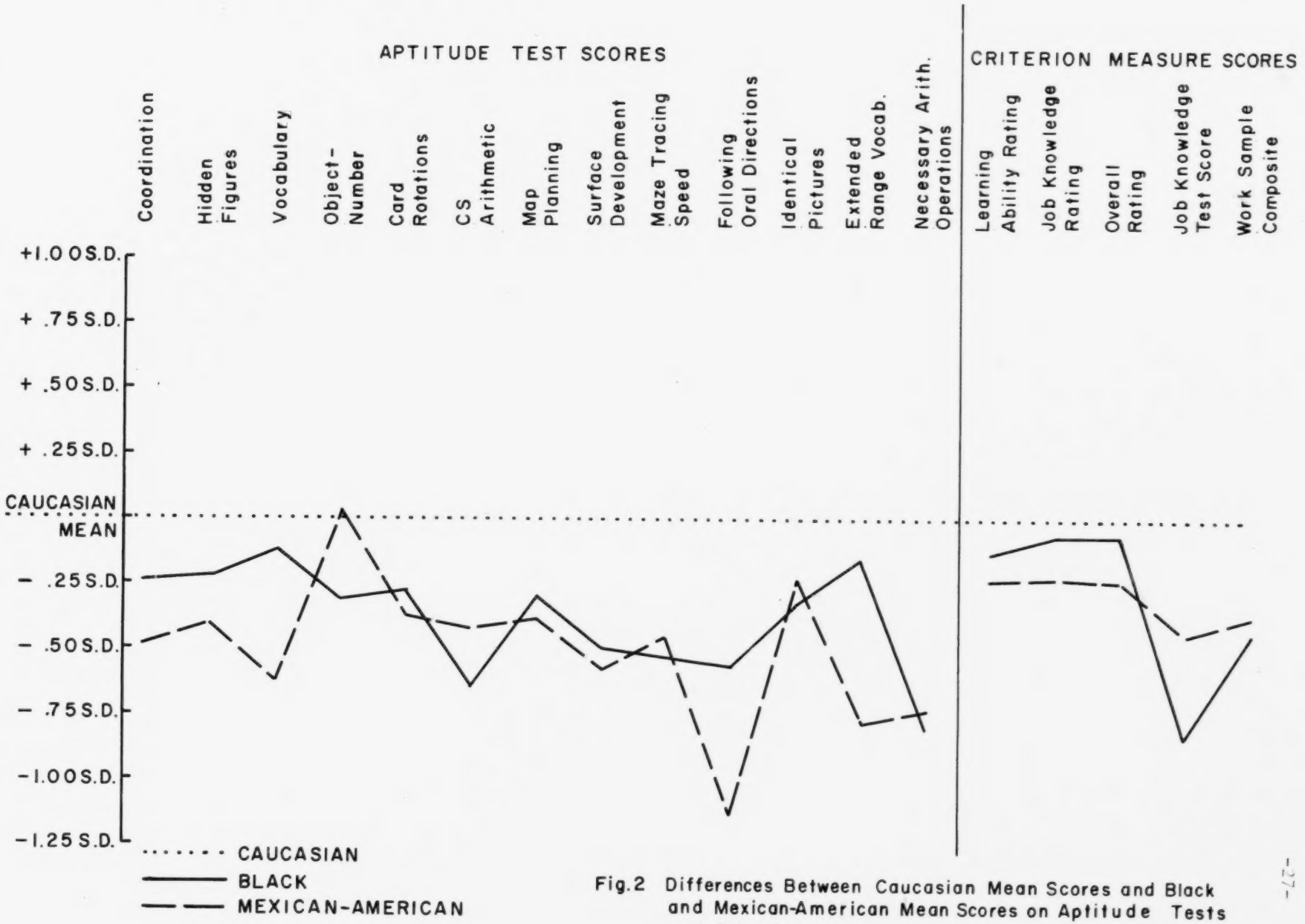


Fig.2 Differences Between Caucasian Mean Scores and Black and Mexican-American Mean Scores on Aptitude Tests and Criterion Measures in Terms of Caucasian Standard Deviation Units CARTOGRAPHIC TECHNICIANS(TOPOCOM)

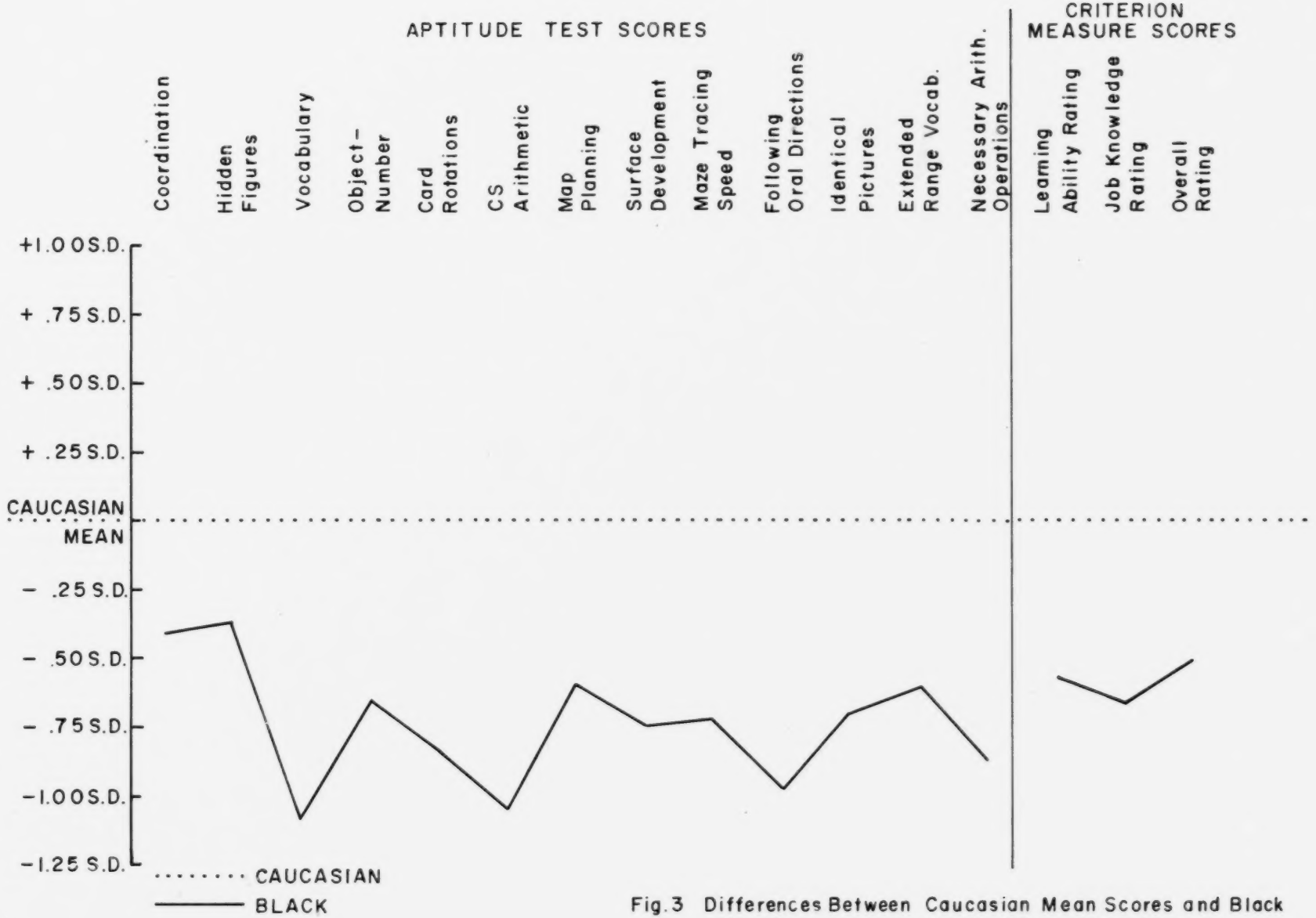


Fig. 3 Differences Between Caucasian Mean Scores and Black Mean Scores on Aptitude Tests and Criterion Measures in Terms of Caucasian Standard Deviation Units CARTOGRAPHIC TECHNICIANS (Coast & Geodetic Survey)

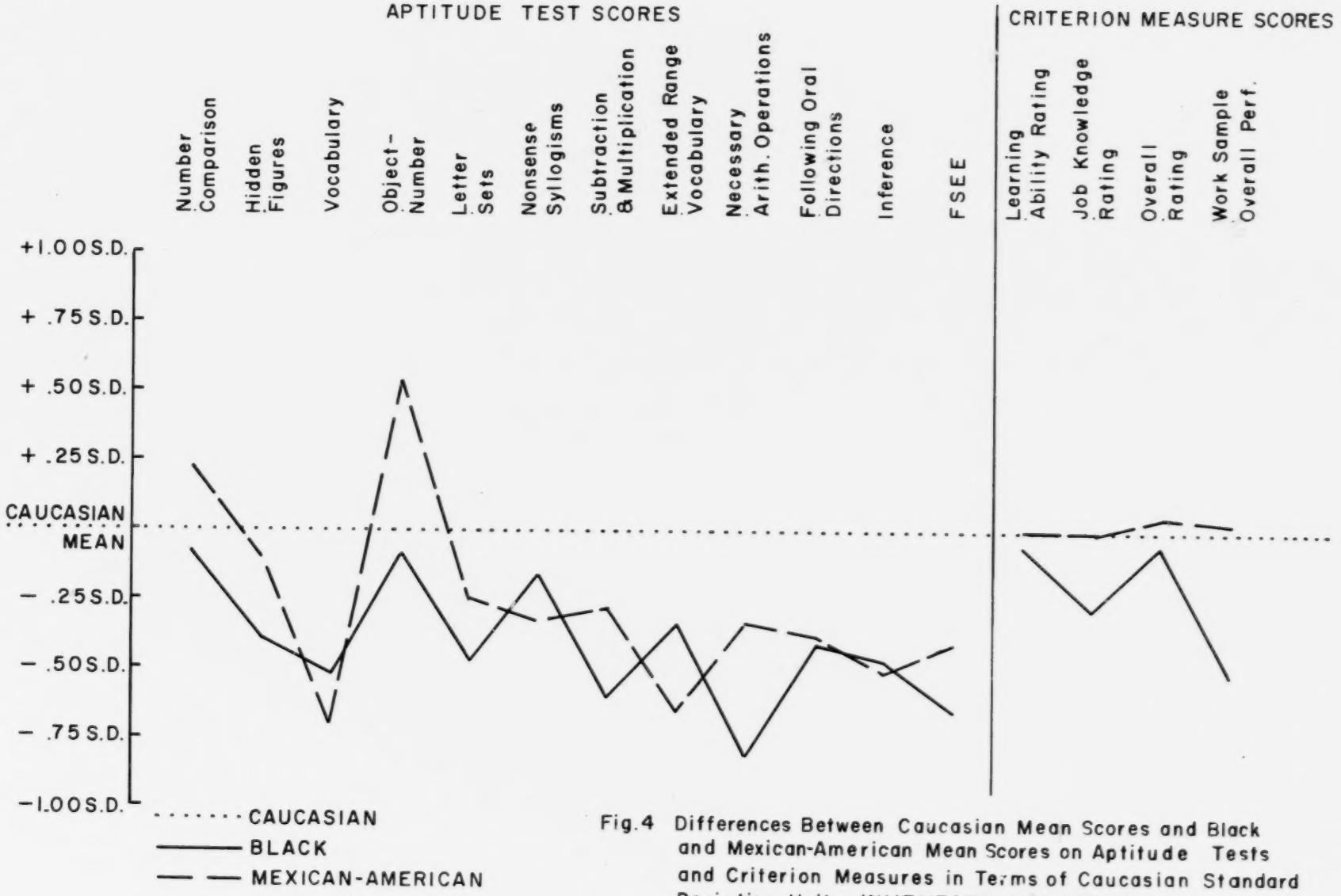


Fig. 4 Differences Between Caucasian Mean Scores and Black and Mexican-American Mean Scores on Aptitude Tests and Criterion Measures in Terms of Caucasian Standard Deviation Units INVENTORY MANAGERS

Table 5

Correlations Between Aptitude Tests and Supervisors' Ratings for Cartographic Technicians by Ethnic Group (TOPOCOM Sample)

	Learning Ability Rating			Overall Performance Rating		
Test	Black N=101	Mexican-American N=99	Caucasian N=240	Black N=101	Mexican-American N=99	Caucasian N=240
Coordination	.15	.17	.21**	.04	.05	.18**
Hidden Figures	.29**	.41**	.25**	.21*	.29**	.21**
Vocabulary	.17	.01	.03	.19	-.02	.01
Object-Number	.21*	.12	.04	.19	.01	.02
Card Rotation	.28**	.19	.31**	.16	.04	.26**
CS Arithmetic	.42**	.34*	.25**	.31**	.21*	.24**
Map Planning	.33**	.39**	.40**	.24**	.23*	.30**
Surface Development	.41**	.35**	.34**	.28**	.21*	.28**
Maze Tracing	.20*	.33**	.32**	.14	.15	.27**
Following Oral Directions	.32**	.32**	.33**	.18	.15	.25**
Identical Pictures	.33**	.26**	.20**	.21*	.18	.14*
Extended Range Vocabulary	.16	.07	-.05	.17	.03	-.07
Necessary Arithmetic Operations	.32**	.36**	.29**	.25**	.22*	.19**

* significant at .05 level
** significant at .01 level

Table 6

Correlations Between Aptitude Tests and Work Sample Overall Score for Inventory Managers by Ethnic Group

Test	Black N=99	Mexican-American N=58	Caucasian N=167
Number Comparison	.17	.36**	.34**
Hidden Figures	.21*	.29*	.30**
Vocabulary	.32**	.41**	.37**
Object-Number	.04	.20	.06
Letter Sets	.28**	.49**	.29**
Nonsense Syllogisms	.29**	.38**	.13
Subtraction & Multiplication	.08	.37**	.13
Extended Range Vocabulary	.28**	.58**	.32**
Necessary Arithmetic Operations	.33**	.60**	.35**
Following Oral Directions	.36**	.41**	.42**
Inference	.39**	.56**	.34**
FSEE (VA + QR)	.37**	.60**	.40**

* significant at .05 level
** significant at .01 level

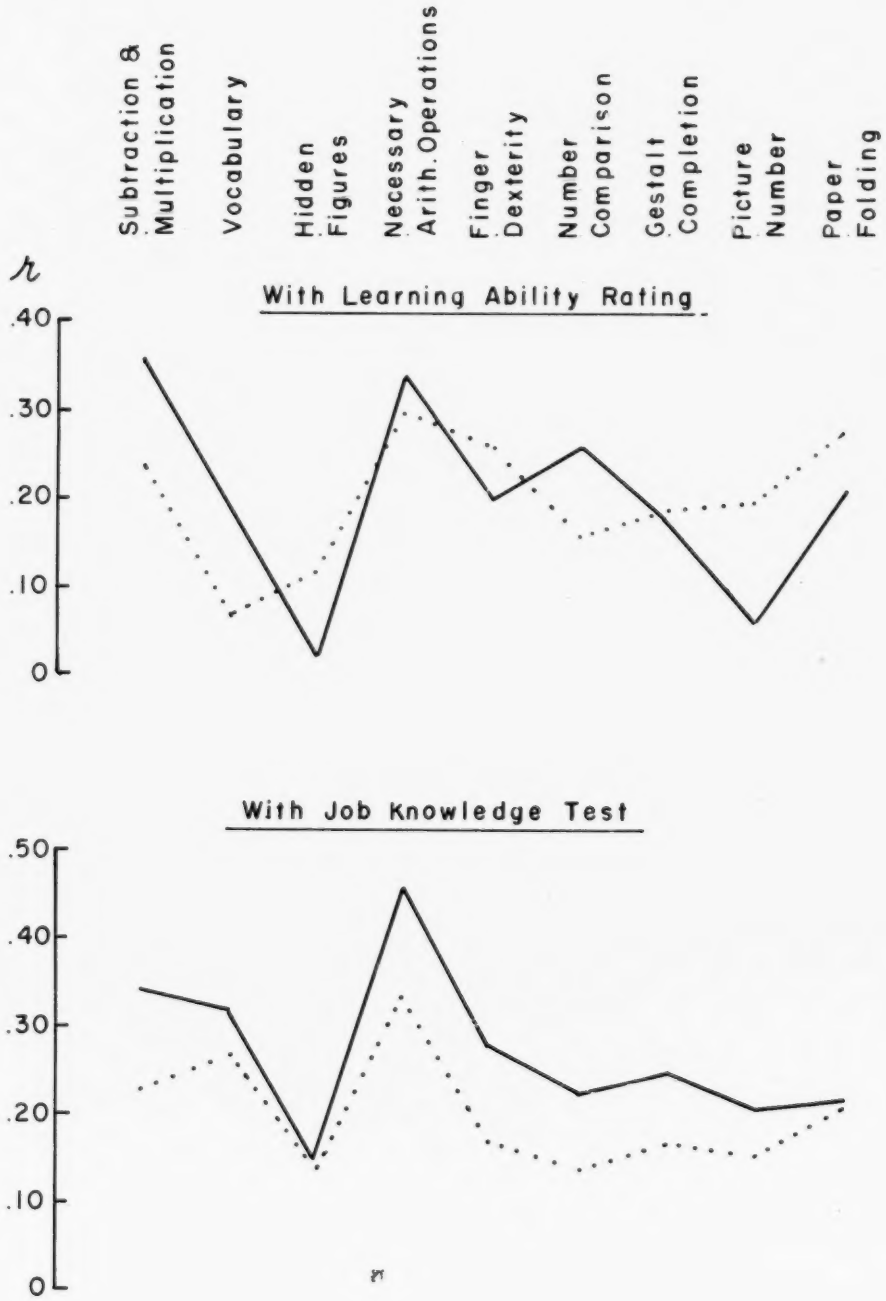


Fig. 5 Validity Coefficients
MEDICAL TECHNICIANS

——BLACK
·····CAUCASIAN

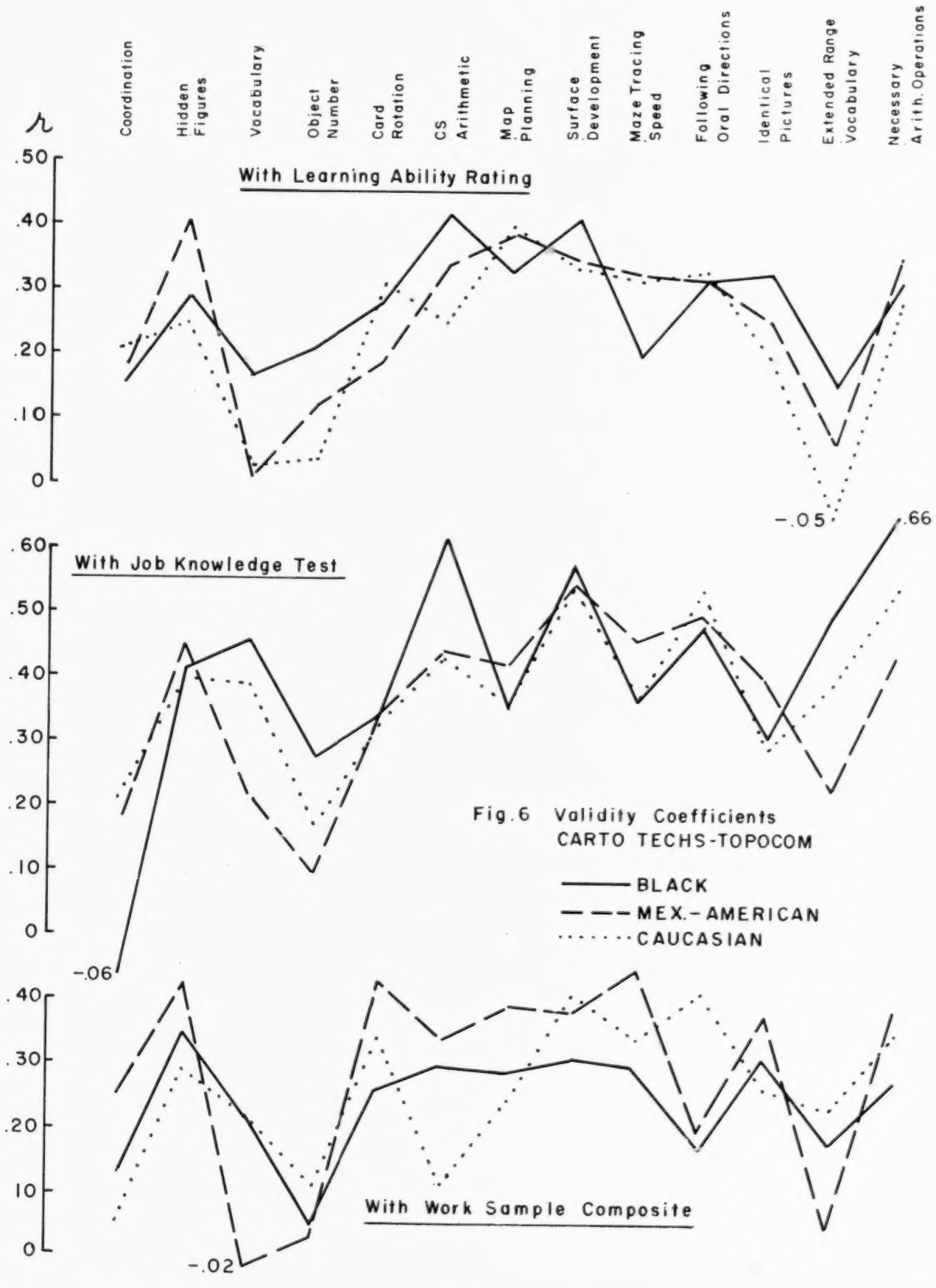


Fig.6 Validity Coefficients
CARTO TECHS-TOPOCOM

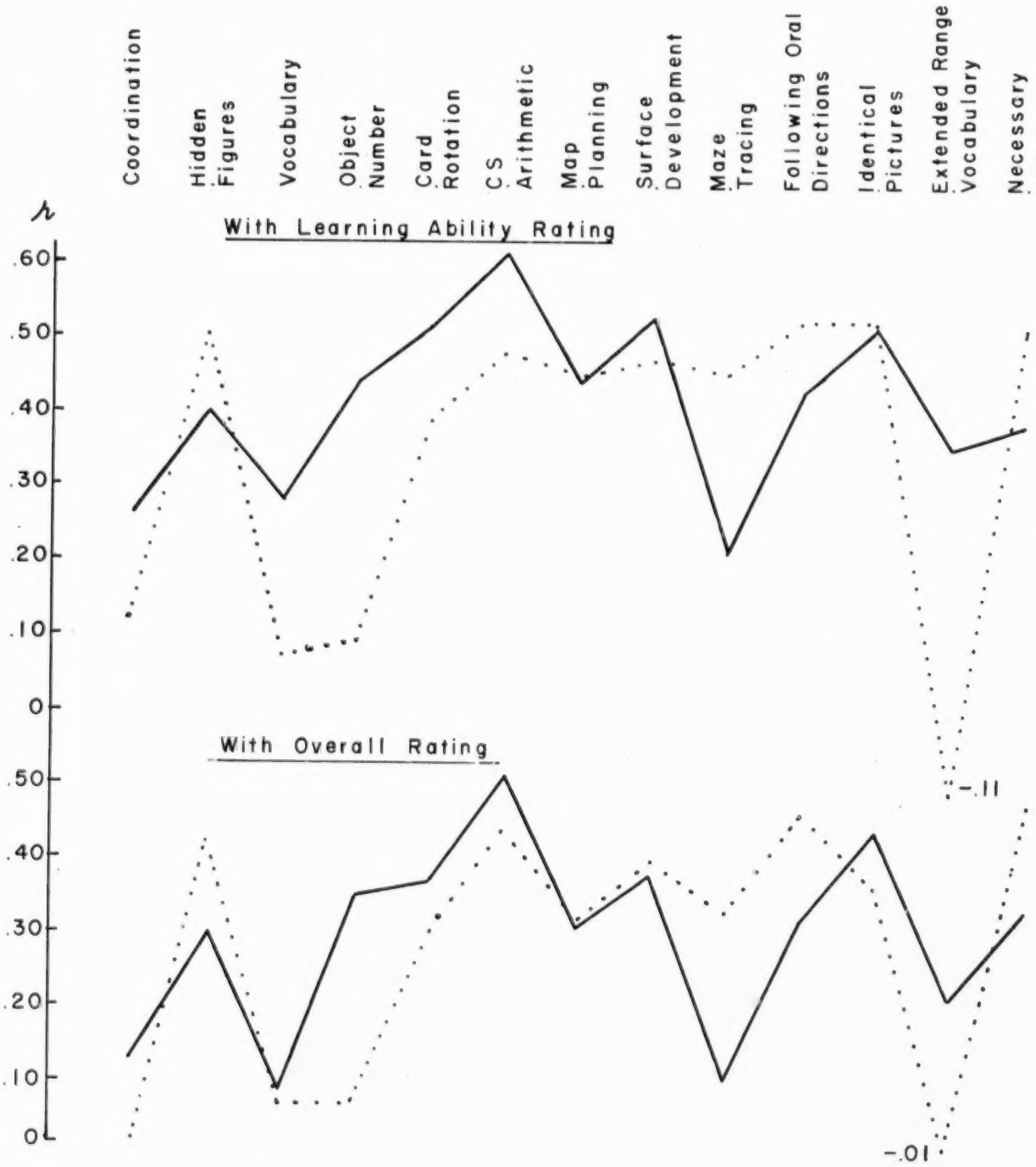


Fig.7 Validity Coefficients
CARTO TECHS
(Coast & Geodetic)

——— BLACK

····· CAUCASIAN

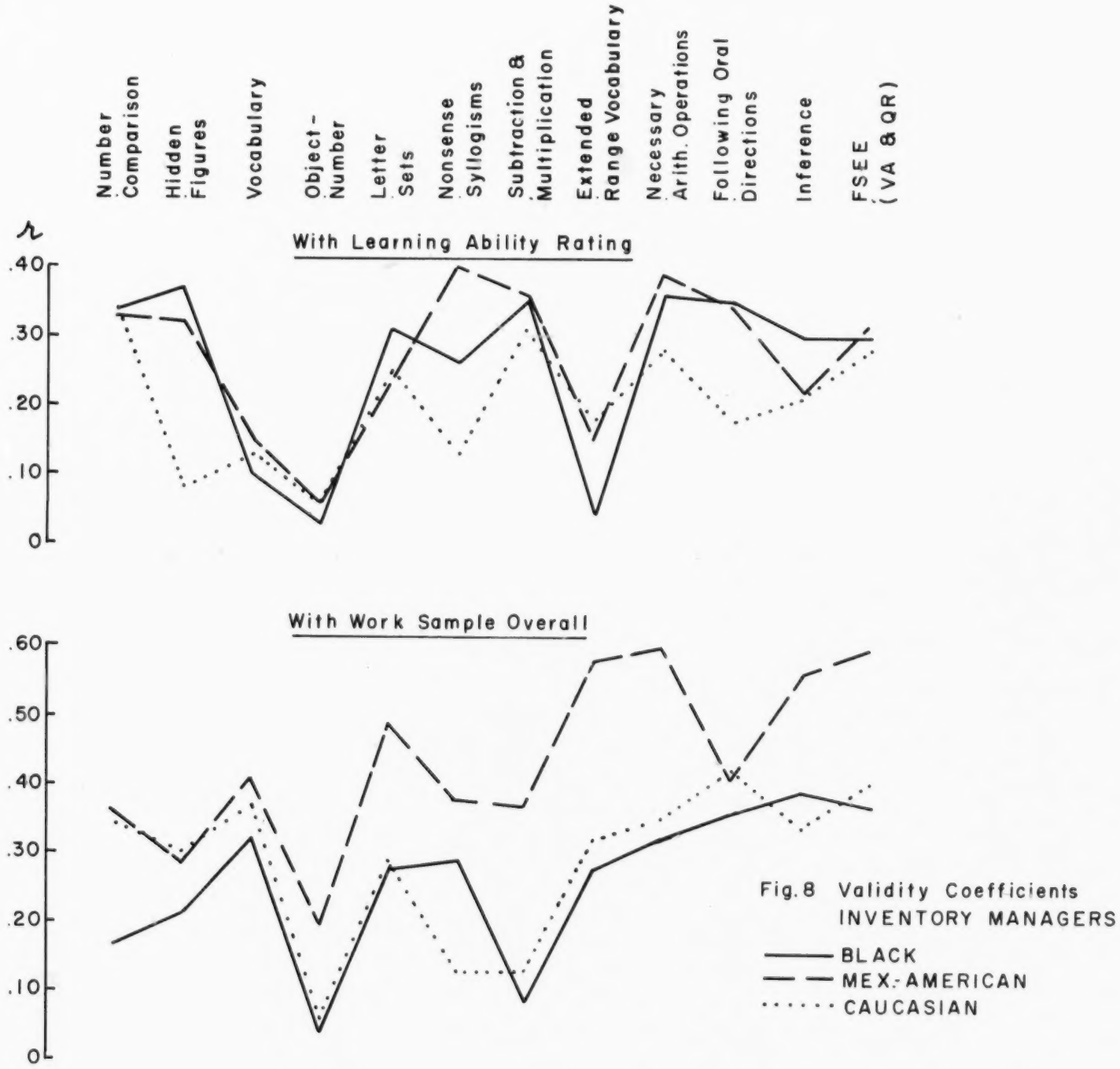


Fig. 8 Validity Coefficients
INVENTORY MANAGERS

Table 7

Comparison of Regression Lines for Different Pairs of Ethnic Groups

Criterion	Occupation	Samples Compared	Number of Regression Lines: No Difference	Location	Difference*
Ratings	Med. Tech.	Black/Caucasian	8	Slope	1
Ratings	Carto. Tech.	Black/Caucasian (C & G)	11	Slope Dispersion	1 1
Ratings	Carto. Tech.	Black/Caucasian (TOPOCOM)	12	Slope	1
Ratings	Carto. Tech.	Mexican-American/ Caucasian (TOPOCOM)	13		0
Ratings	Inv. Mgr.	Black/Caucasian	12		0
Ratings	Inv. Mgr.	Mexican-American/ Caucasian	12		0
Job Knowledge Test	Med. Tech.	Black/Caucasian	0	Slope Intercept	1 8
Job Knowledge Test	Carto. Tech.	Black/Caucasian (TOPOCOM)	4	Intercept	9
Job Knowledge Test	Carto. Tech.	Mexican-American/ Caucasian (TOPOCOM)	0	Dispersion Intercept	1 12
Work Sample	Carto. Tech.	Black/Caucasian (TOPOCOM)	3	Intercept Slope	9 1
Work Sample	Carto. Tech.	Mexican-American/ Caucasian (TOPOCOM)	0	Dispersion Slope Intercept	7 2 4
Work Sample	Inv. Mgr.	Black/Caucasian	0	Dispersion	12
Work Sample	Inv. Mgr.	Mexican-American/ Caucasian	9	Dispersion	3

* Significant at .05 level or better.

Table 8

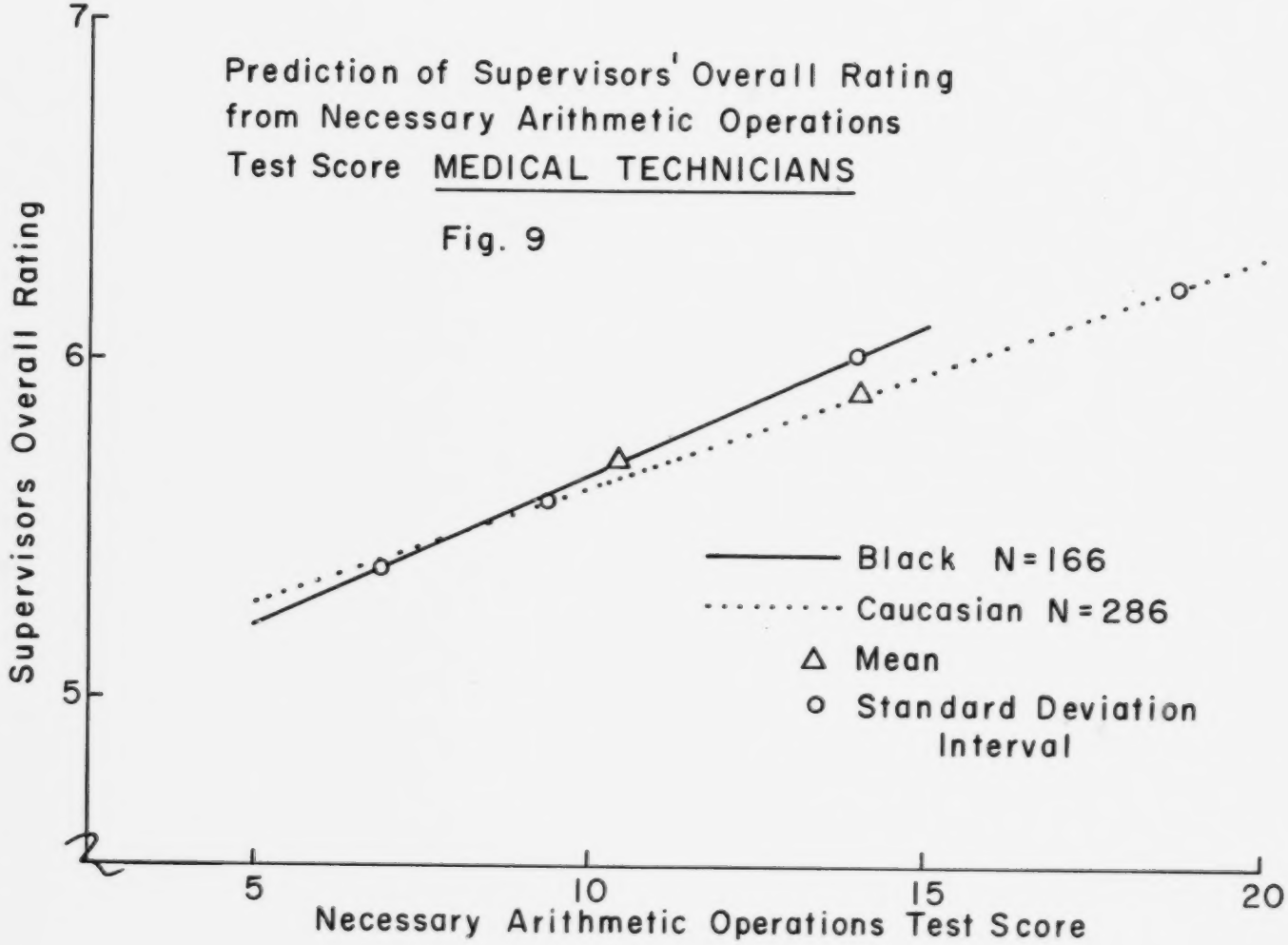
Comparison of Regression Lines for Inventory Managers

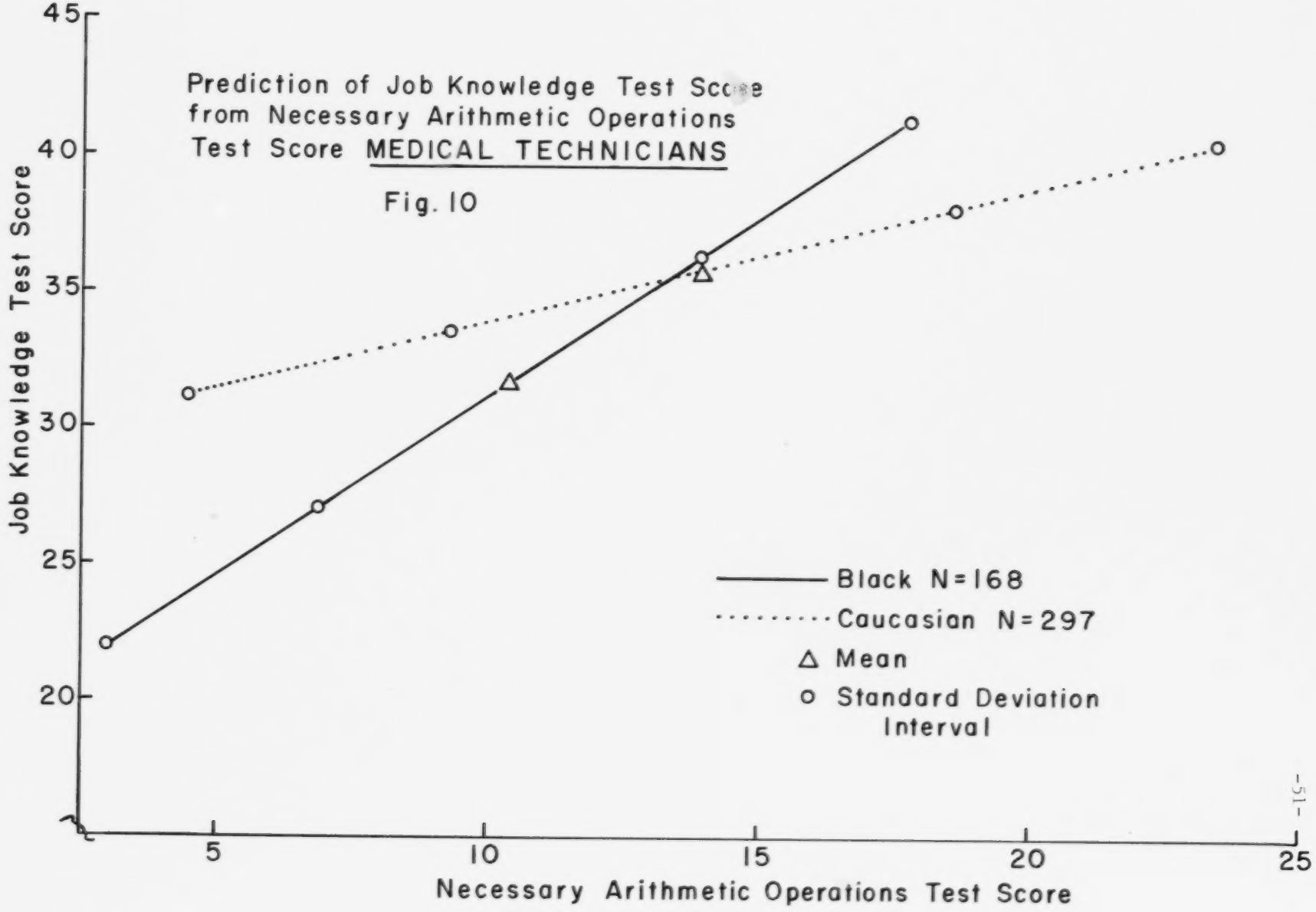
Criterion	Occupation	Samples Compared	Number of Regression Lines: No Difference	Location	Difference*
Ratings	Inventory Manager	Black and Caucasian at Detroit, Phila., and Dayton Installations	10	Intercept	2
Ratings	Inventory Managers	Mexican-American and Caucasian at San Antonio	12		0
Work Sample	Inventory Manager	Black and Caucasian at Detroit, Phila., and Dayton Installations	0	Dispersion	12
Work Sample	Inventory Manager	Mexican-American and Caucasian at San Antonio	12		0

* Significant at .05 level or better.

Prediction of Supervisors' Overall Rating
from Necessary Arithmetic Operations
Test Score MEDICAL TECHNICIANS

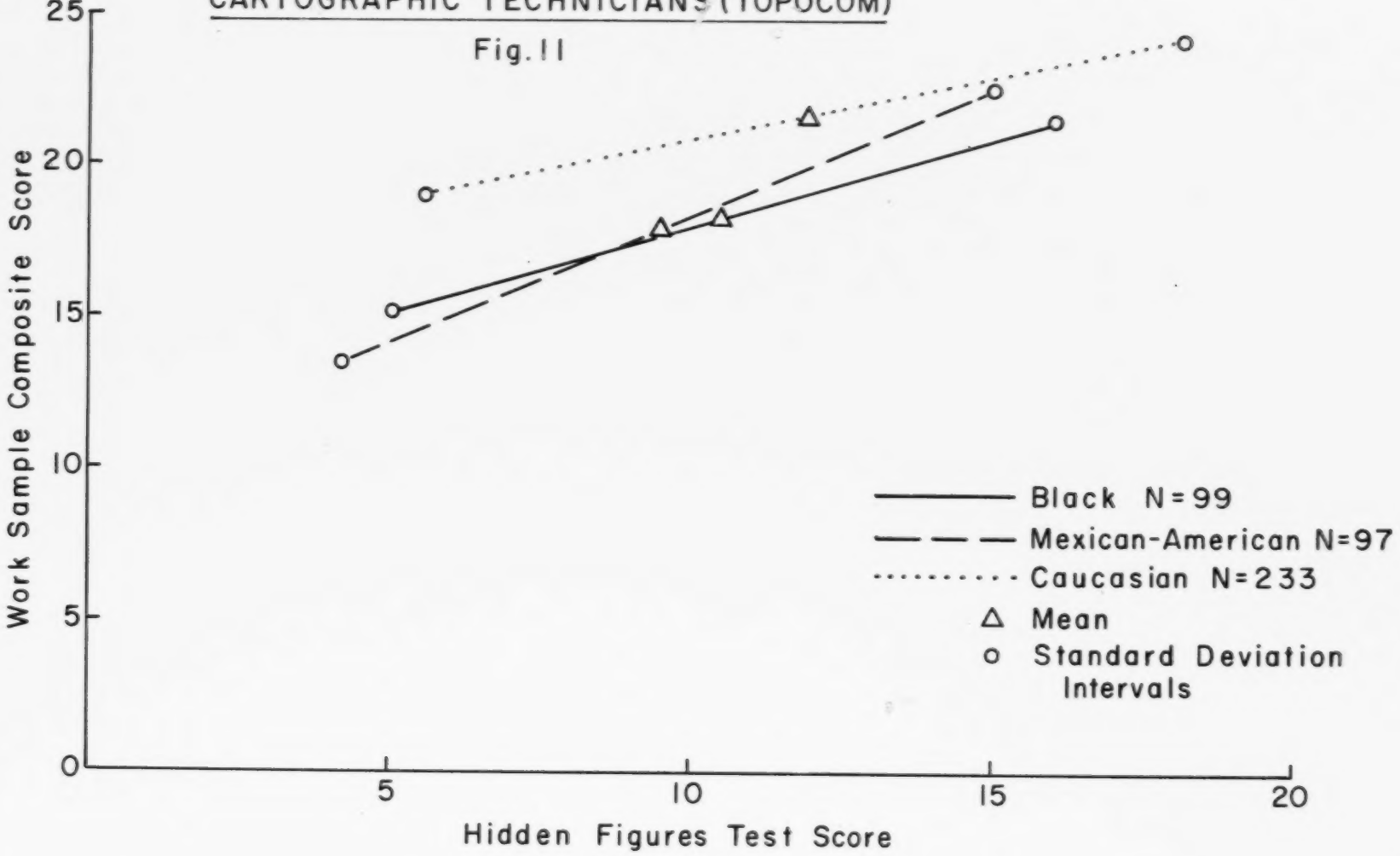
Fig. 9




Prediction of Job Knowledge Test Score
from Necessary Arithmetic Operations
Test Score MEDICAL TECHNICIANS
Fig. 10
Job Knowledge Test Score
45
40
35
30
25
20
5
10
15
20
25
Necessary Arithmetic Operations Test Score
Black N=168
Caucasian N=297
Mean
Standard Deviation Interval

Prediction of Work Sample Composite Score
from Hidden Figures Test Score
CARTOGRAPHIC TECHNICIANS (TOPOCOM)

Fig. 11



Prediction of Supervisors' Overall Rating
from Subtraction & Multiplication Test Score

INVENTORY MANAGERS

Fig. 12

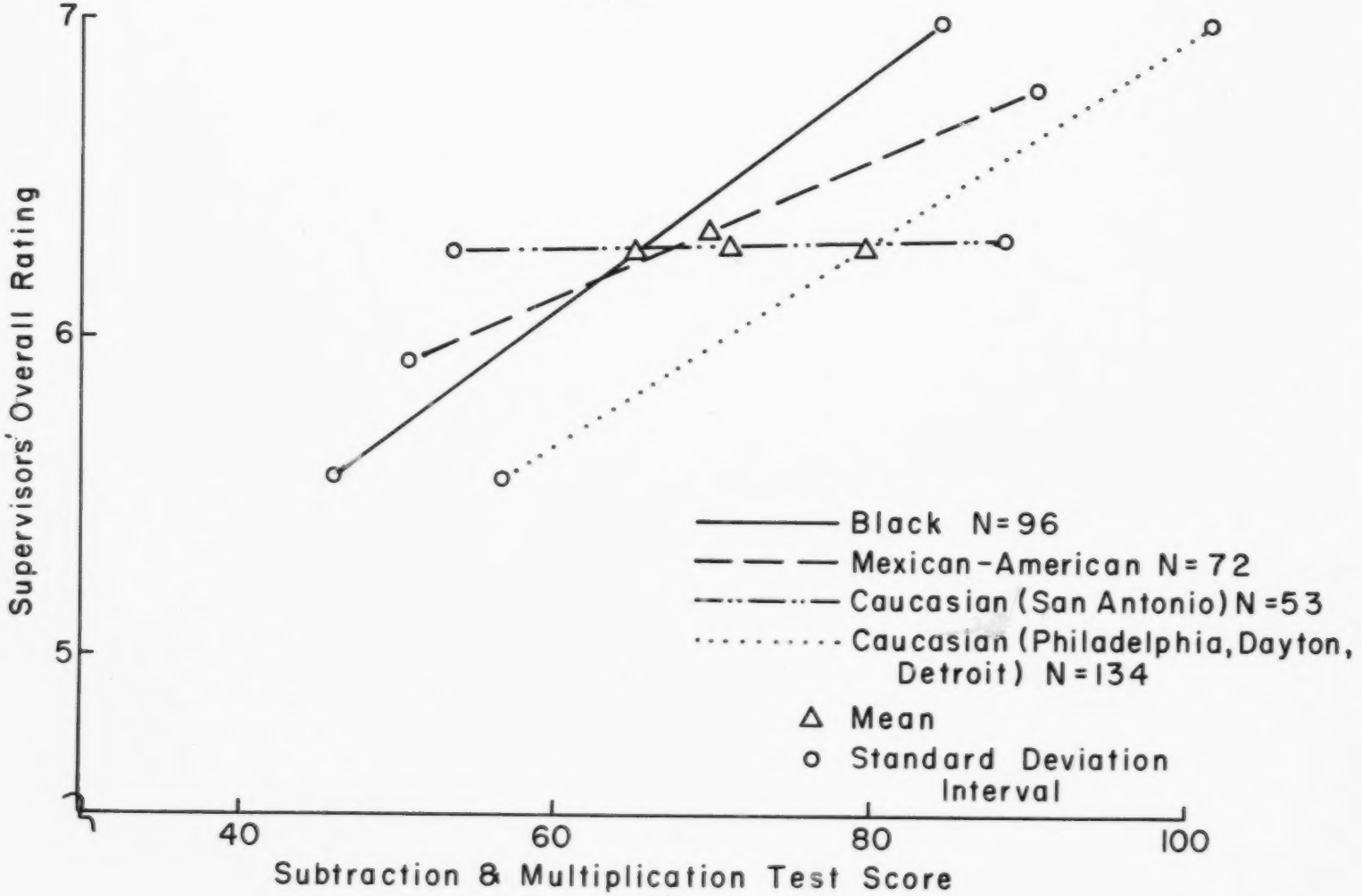


Table 9

CRITERION SCORE MEANS FOR CARTOGRAPHIC TECHNICIANS

AT DIFFERENT SCORE LEVELS ON THE

MAP PLANNING TEST

Test Score	Mean Supervisors' Overall Rating			Mean Job Knowledge Test Score			Mean Work Sample Composite		
	Black	Mexican-American	Caucasian	Black	Mexican-American	Caucasian	Black	Mexican-American	Caucasian
24 +	6.0 N=18	6.7 N=16	6.6 N=56	44.3 N=18	38.9 N=16	45.3 N=56	21.9 N=18	24.8 N=16	24.2 N=56
17 - 23.9	5.8 N=28	5.9 N=34	6.0 N=101	34.0 N=28	30.7 N=34	41.4 N=101	19.0 N=28	20.4 N=33	22.0 N=98
10 - 16.9	5.1 N=45	5.5 N=35	5.2 N=67	35.1 N=45	29.9 N=35	37.2 N=67	17.2 N=43	14.4 N=34	20.0 N=65
- 9.9	5.0 N=10	5.2 N=16	5.2 N=17	24.7 N=10	22.4 N=14	29.7 N=17	13.7 N=10	13.1 N=16	16.0 N=17

Table 10

CRITERION SCORE MEANS FOR INVENTORY MANAGEMENT SPECIALISTS AT DIFFERENT SCORE LEVELS ON THE SUBTRACTION AND MULTIPLICATION TEST

Test Scores	Mean Supervisors' Overall Rating			Mean Work Sample Overall Rating		
	Black	Mexican-American	Caucasian	Black	Mexican-American	Caucasian
90+	7.2 N=14	7.1 N=14	6.9 N=54	6.2 N=14	9.1 N=10	9.4 N=43
70 - 89	6.3 N=28	6.4 N=22	6.4 N=65	6.7 N=24	11.3 N=18	8.4 N=56
50 - 69	6.4 N=48	6.3 N=23	5.8 N=48	7.1 N=44	8.4 N=18	8.6 N=40
- 49	4.8 N=22	5.6 N=13	5.3 N=20	5.7 N=18	5.9 N=11	8.3 N=18

Table 11

COMPARISONS OF CRITERION SCORES

PREDICTED FOR BLACK JOB INCUMBENTS

FROM MULTIPLE REGRESSION EQUATIONS

AT DIFFERENT SCORE LEVELS

Score level	Higher predicted score using weights developed on Black sample	Equal predicted score	Higher predicted score using weights developed on Caucasian sample
One standard deviation above mean	8	0	4
At mean	5	0	8
One standard deviation below mean	3	0	10

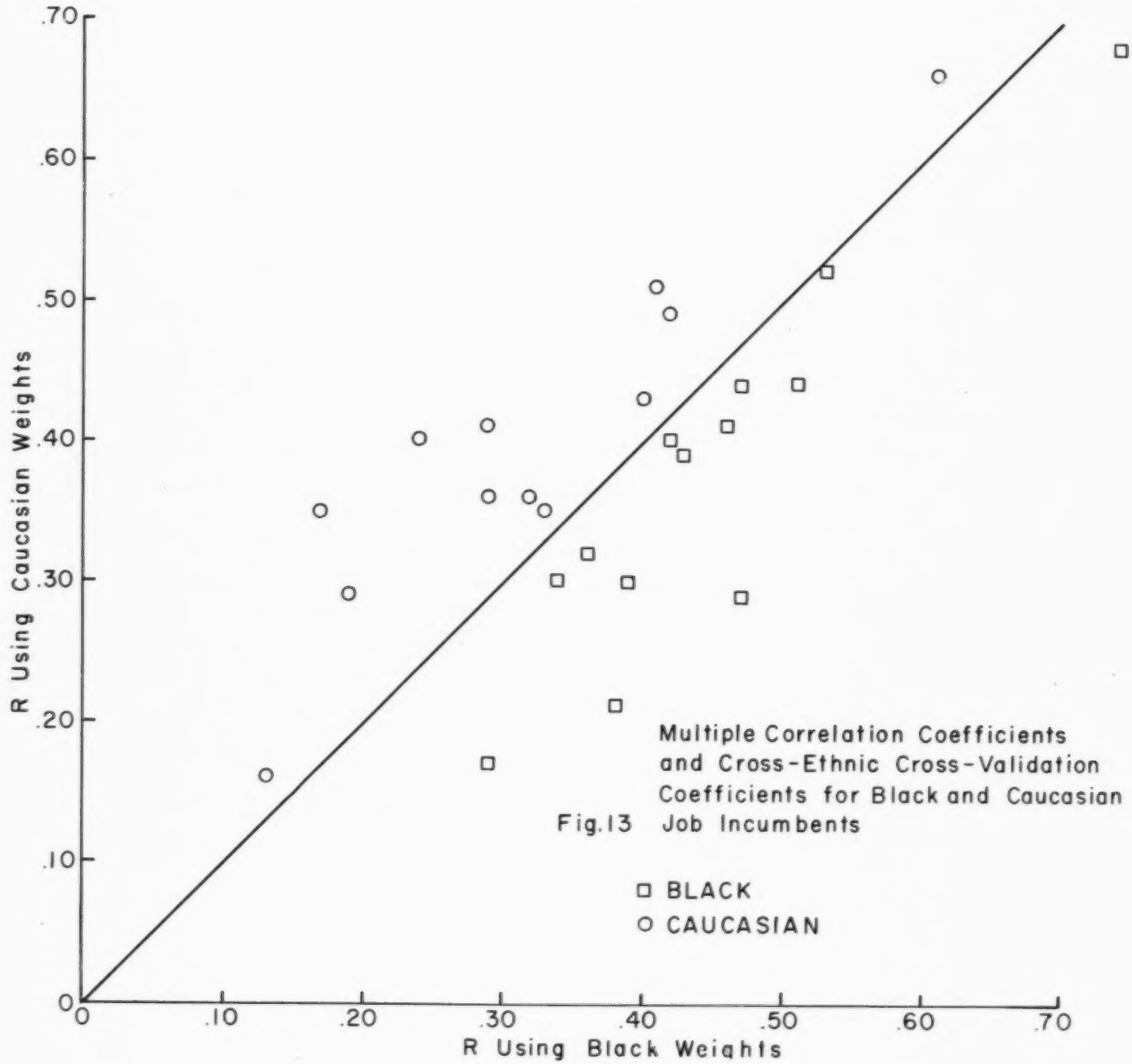
COMPARISONS OF CRITERION SCORES

PREDICTED FOR MEXICAN-AMERICAN JOB INCUMBENTS

FROM MULTIPLE REGRESSION EQUATIONS

AT DIFFERENT SCORE LEVELS

Score level	Higher predicted score using weights developed on Mexican-American sample	Equal predicted score	Higher predicted score using weights developed on Caucasian sample
One standard deviation above mean	7	1	1
At mean	8	0	1
One standard deviation below mean	4	1	4



Multiple Correlation Coefficients and Cross-Ethnic Cross-Validation Coefficients for Black and Caucasian Job Incumbents

Fig.13

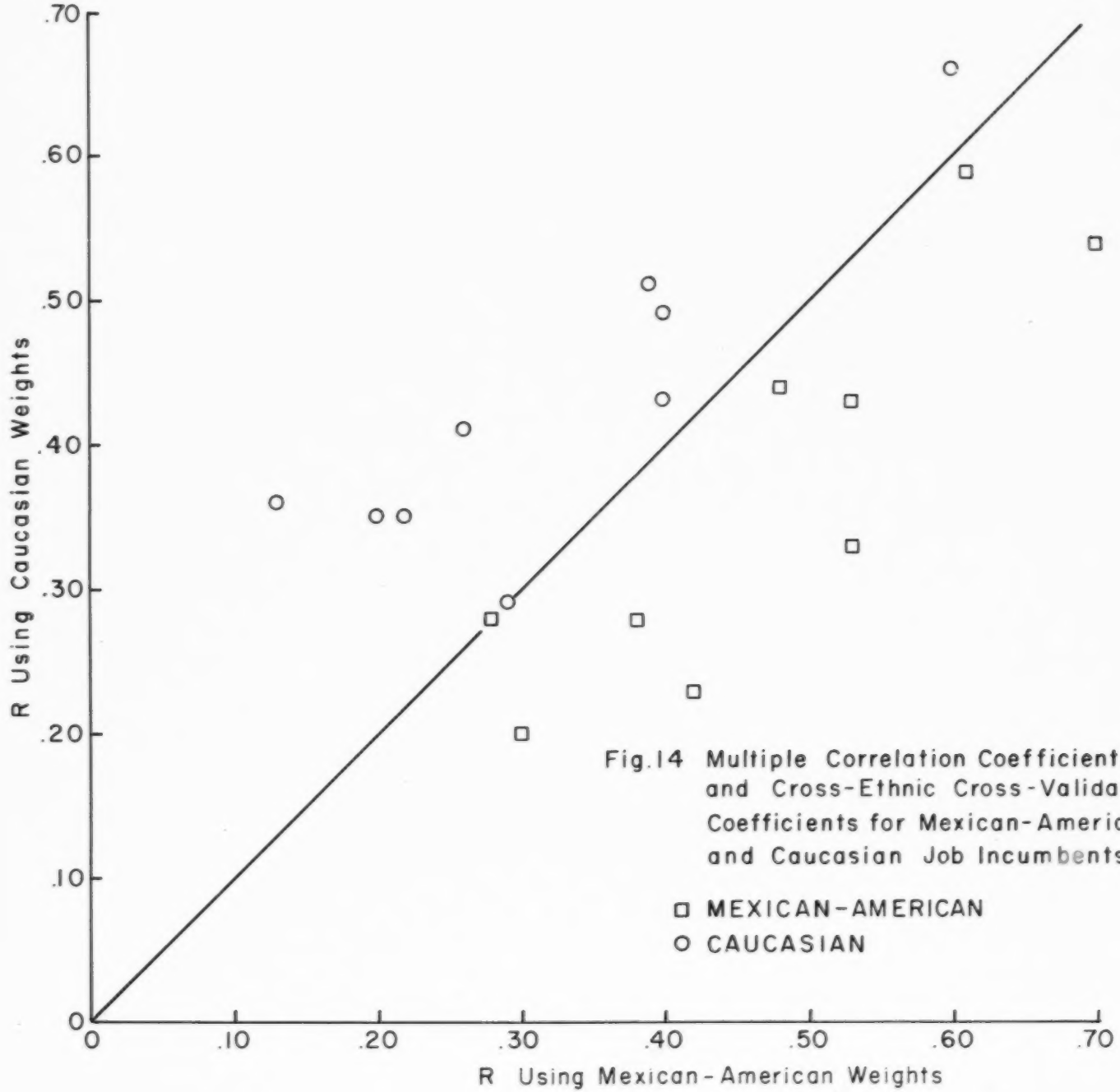


Fig.14 Multiple Correlation Coefficients and Cross-Ethnic Cross-Validation Coefficients for Mexican-American and Caucasian Job Incumbents

Table 12

SUPERVISORS' RATINGS BY RACE OF RATER AND RACE OF RATEE

Medical Technicians

RACE OF TECHNICIAN: RACE OF RATER:	Black Black N = 43-50	Caucasian Black N = 146-148	Black Caucasian N = 38-41	Caucasian Caucasian N = 272-276
Flexibility	6.2	5.5	5.3	5.4
Organization	6.3	6.1	5.6	6.0
Interest	6.1	6.2	5.4	5.8
Learning Ability	6.6	5.8	5.6	6.0
Job Knowledge	6.2	5.5	5.1	5.3
Technique	6.5	5.8	5.8	6.0
Need for Supervision	6.5	5.9	5.6	6.1
Communication	6.3	6.0	5.3	5.7
Overall Rating	6.3	5.8	5.6	5.9

Table 13

MEAN RATINGS BY RACE OF RATER

	Black raters			Caucasian raters		
	Higher mean ratings to Blacks	Equal	Higher mean ratings to Caucasians	Higher mean ratings to Caucasians	Equal	Higher mean ratings to Blacks
	Number of scales			Number of scales		
Medical Technicians	8	0	1	9	0	0
Cartographic Technicians (TOPOCOM)	7	0	1	8	0	0
Cartographic Technicians (C & G)	2	1	5	6	0	2
Inventory Management Specialists	6	1	3	7	2	1
Total	23	2	10	30	2	3

	Mexican-American raters			Caucasian raters		
	Higher mean ratings to Mexican-Americans	Equal	Higher mean ratings to Caucasians	Higher mean ratings to Caucasians	Equal	Higher mean ratings to Mexican-Americans
	Number of scales			Number of scales		
Cartographic Technicians (TOPOCOM)	8	0	0	4	3	1
Inventory Management Specialists				6	2	2
Total	8	0	0	10	5	3

Table 14

CORRELATION OF LEARNING ABILITY RATINGS WITH APTITUDE TESTS, JOB KNOWLEDGE TEST, AND WORK SAMPLES

BY RACE OF RATER AND RACE OF RATEE

Cartographic Technicians (TOPOCOM Samples)

RACE OF TECHNICIAN: RACE OF RATER:	Black Black	Caucasian Black	Mexican-American Mexican-American	Caucasian Mexican-American	Black Caucasian	Mexican-American Caucasian	Caucasian Caucasian
	N=20	N=53	N=97	N=26	N=99	N=99	N=240
Coordination	.47	-.04	.15	.07	.11	.14	.21
Hidden Figures	.33	.32	.37	.36	.22	.35	.21
Vocabulary	.48	.06	.03	.21	.13	.03	.02
Object-Number	.42	.08	.03	.46	.14	.14	.04
Card Rotation	.42	.35	.27	.44	.21	.08	.29
Arithmetic	.59	.20	.31	.36	.37	.26	.22
Map Planning	.51	.33	.35	.53	.28	.29	.37
Surface Development	.36	.26	.31	.61	.36	.27	.32
Maze Tracing	.51	.32	.33	.57	.16	.22	.29
Following Oral Directions	.44	.19	.24	.51	.28	.31	.32
Identical Pictures	.34	.26	.26	.40	.31	.19	.17
Extended Range Vocabulary	.22	-.09	.05	.09	.12	.13	-.08
Necessary Arithmetic Operations	.55	.23	.33	.46	.26	.31	.29
Job Knowledge	.48	.55	.42	.66	.30	.46	.38
Geometric Restitution	.11	.12	.19	.48	.11	.22	.20
Logical Contouring	.43	-.06	.33	.38	.13	.24	.11
Pull-Up	.67	.41	.44	.41	.15	.33	.24

Table 15

CORRELATIONS OF OBJECTIVE MEASURES WITH LEARNING ABILITY RATINGS

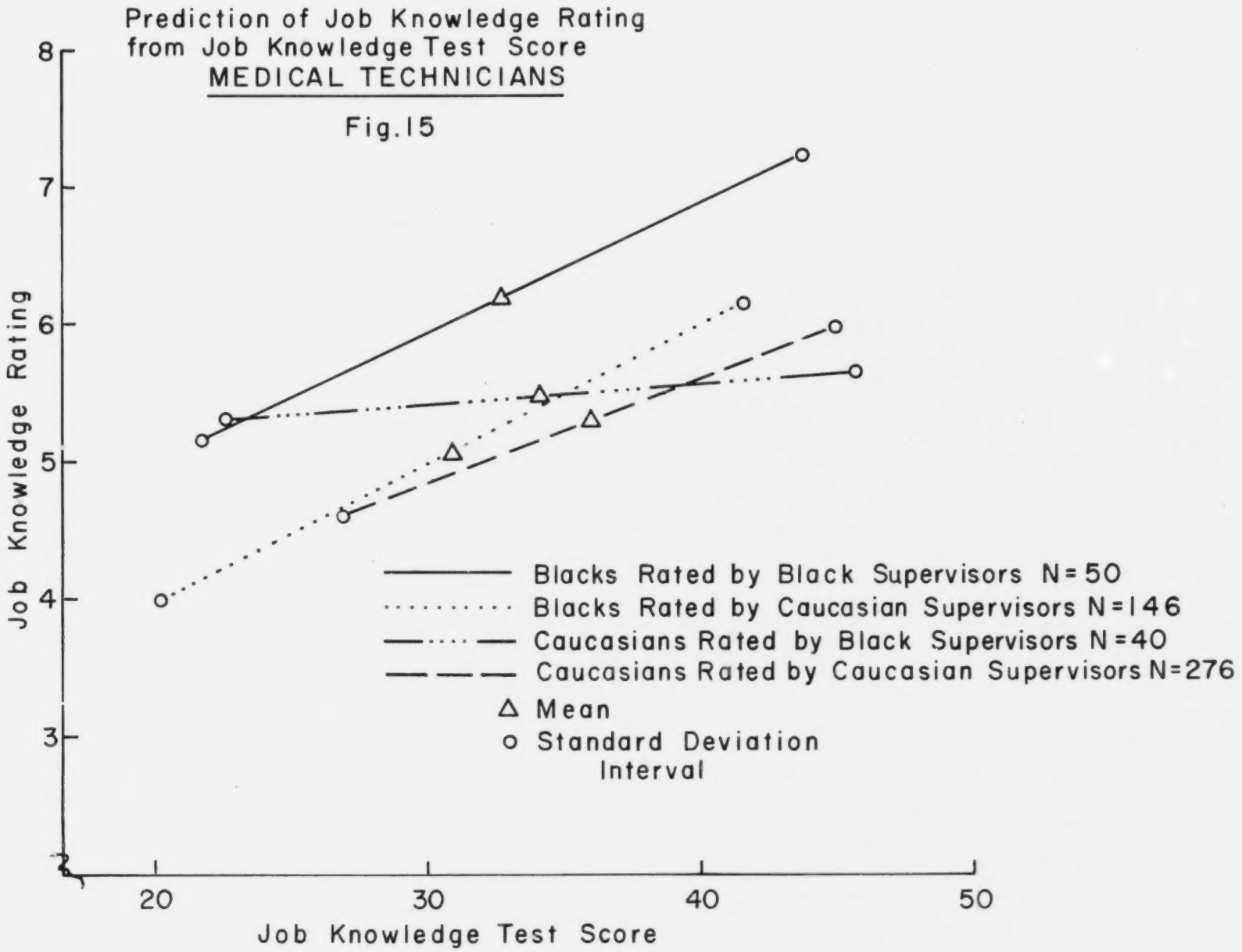
	Black raters			Caucasian raters		
	Higher validities for Blacks	Equal	Higher validities for Caucasians	Higher validities for Caucasians	Equal	Higher validities for Blacks
Medical Technicians	5	1	4	5	0	5
Cartographic Technicians (TOPOCOM)	15	0	2	9	0	8
Cartographic Technicians (C & G)	7	0	6	7	1	5
Inventory Management Specialists	11	0	4	2	1	12
Total	38	1	16	23	2	30

	Mexican-American raters			Caucasian raters		
	Higher validities for Mexican-Americans	Equal	Higher validities for Caucasians	Higher validities for Caucasians	Equal	Higher validities for Mexican-Americans
Cartographic Technicians (TOPOCOM)	3	0	14	6	0	11
Inventory Management Specialists				1	1	13
Total	3	0	14	7	1	24

Table 16

Average Correlation of Learning Ability Rating With Objective Measures by Race of Rater and Race of Ratee

Occupation	Ethnic group of Ratee	Ethnic group of rater: Black	Mexican-American	Caucasian
Medical Technician	Black	.27		.25
	Caucasian	.17		.23
Cartographic Technician (TOPOCOM)	Black	.44		.22
	Mexican-American		.26	.24
	Caucasian	.22	.42	.22
Cartographic Technician (C & G)	Black	.59		.27
	Caucasian	.47		.27
Inventory Manager	Black	.40		.26
	Mexican-American			.28
	Caucasian	.28		.20
Total	Black	.44		.25
	Mexican-American		.26	.26
	Caucasian	.28	.42	.25


Prediction of Job Knowledge Rating
from Job Knowledge Test Score
MEDICAL TECHNICIANS
Fig.15
Job Knowledge Rating
8
7
6
5
4
3
2
20
30
40
50
Job Knowledge Test Score
Blacks Rated by Black Supervisors N=50
Blacks Rated by Caucasian Supervisors N=146
Caucasians Rated by Black Supervisors N=40
Caucasians Rated by Caucasian Supervisors N=276
Mean
Standard Deviation Interval

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE:

TECHNICAL CRITIQUE

Anne Anastasi

Professor of Psychology

Fordham University

To give a systematic technical critique of a study of such vast scope is obviously impossible within the time available. It is fortunate, therefore, that the general experimental design and the procedures for data gathering and analysis are such that they can be simply characterized as representing a high level of technical excellence. The study is in many ways a model for the validation of personnel selection tests. Against this background, I have chosen four questions for brief discussion. Some concern specifically procedures or results of the present study; others take that study as a point of departure for a consideration of broader methodological issues.

Validity studies of incumbents

The first is the familiar question regarding the use of job applicants or of present employees in validation studies. The ideal procedure would be to test a large sample of job applicants, hire them all, and follow them up until a satisfactory criterion measure of job performance becomes available on each. For many reasons, this procedure is impracticable, except in rare situations. What, then, are some of the major implications of utilizing incumbents as was done in the present study?

One likely characteristic of an incumbent sample, as compared to an applicant sample, is a restriction of range in job-relevant variables because of preselection. Insofar as this occurs, its effect is to lower validity coefficients of predictors. Preselection, of course, can operate either at the time of employment or subsequently through discharge or voluntary dropout. For the

present study, the report does state that little preselection on tests occurred for Medical Technicians and Cartographic Technicians. Further information would be desirable, however, especially with regard to possible ethnic differences in extent of preselection.

Another possible selective factor stems from the employees' option not to participate in the study. The literature on volunteer error reveals a number of systematic differences between participants and nonparticipants under such circumstances. It appears, however, that there were relatively few refusals to participate in this project, probably because of the effective advance communication regarding the purpose and nature of the study. Refusals were more frequent among the Inventory Management Specialists than in the other two occupational groups; but the attrition from this source seems to have been fairly uniform across ethnic groups.

Incumbents also tend to differ from job applicants in their test-taking motivation. When taking a test for research purposes only, with the assurance that the scores can in no way affect their job status, individuals may not respond as they would when tested for selection purposes. In the genuine selection situation, some persons may try harder and perform better on aptitude tests; others may become overanxious and perform more poorly. Furthermore, some tests would probably be more susceptible than others to these attitudinal differences. It is thus impossible to generalize about the likely effects of these differences in test-taking motivation. Much depends, too, on the prior communication about the project, the rapport established by the examiner, and the cooperation elicited from the examinees. In the present study, these procedural matters seem to have been handled with unusual care.

Still another implication of the use of incumbents pertains to the possible influence of job experience on both predictor and criterion scores. In

this connection, it is desirable to have data on any ethnic differences in length of time on the job. Any background data that could help us to understand why ethnic groups perform differently on both predictor and criterion measures would represent a significant contribution to knowledge. We need as much information as we can find on how test scores are related to the individual's reactional biography. Insofar as individual differences in job tenure are appreciable, however, it would also be interesting to have the correlation between this variable and both predictor and criterion scores.

The criterion in the validation of personnel tests

My second major topic centers around the crucial importance of the criterion. Insofar as predictors are evaluated on the basis of their relation to criterion measures, a validation study can be no better than the quality of its criterion data. Yet, in real-life situations, good criterion data are hard to come by.

There are many possible sources of criterion data and the optimum choice certainly differs with the nature of the job. Because any one type of criterion measure is likely to have some deficiencies, however, a combination of diverse measures seems to be indicated in practically all situations. The present study utilized three quite different types, including a work sample test, a job knowledge test, and ratings of both overall job performance and specific behavioral characteristics.

At first sight it might seem that the most realistic criterion measures are those based on actual job performance over a designated minimum time period. Such indices, however, present serious practical difficulties. In many jobs, there are no objective output records, and none may be feasible. Moreover, the conditions under which individual workers carry out their job functions may vary so much as to introduce excessive error variance into objective output records.

The closest approximation to job performance records, that at the same time provides uniform working conditions, is the standardized work sample test. The coverage of job functions in such a test can be checked directly against job analysis data to assess its content validity, much as is done for educational achievement tests. In fact, the use of work sample tests as criterion measures is similar to the validation of scholastic aptitude tests against the students' subsequent performance on standardized achievement tests. It would also be desirable to provide reliability data, including not only rater reliability as was evidently done in the present study, but also, if possible, some sort of parallel form reliability.

For many kinds of jobs, a paper-and-pencil test of factual job knowledge is clearly appropriate. Such a test provides a good supplement for the work sample test. In some cases, it may have to serve as a substitute for work samples, because of inadequate time or facilities. In both their development and administration, work sample tests are very time consuming. A possible danger in the exclusive reliance on job knowledge tests is that they may demand a higher level of reading comprehension or verbal ability than is required by the job. This is by no means an insoluble problem, however. With proper item formulation and with procedural adaptations, such tests could be administered to illiterates, foreign-speaking persons, or groups with other special testing needs. In the present study, there is some evidence that the performance of the Mexican-American Cartographic Technicians on the Job Knowledge Test may have been somewhat poorer than on the Work Sample because of language handicap.

As in the case of work samples, the content validity of a job knowledge test can be checked against job analysis data. Some measure of reliability, such as a coefficient of internal consistency, is also desirable.

Ratings are commonly used in industrial validation studies for a variety

of reasons. They are often already available through routine personnel procedures; they require no worker time, as do tests; and they represent an index of actual job performance, in which the supervisor can take into account variations in working conditions and other uncontrolled factors and presumably make some adjustment for them. On the other hand, ratings are subject to many well-known random, as well as constant, errors of judgment. Because extensive data on ratings are provided by the present study, I shall discuss the use of ratings as criterion measures as a separate topic, the third in my list.

Ratings as criterion measures

The present study provides considerable evidence that ratings may not be a satisfactory criterion measure, especially in validation studies across ethnic groups. First, the intercorrelations of the ratings for different traits reveal a pronounced halo effect. Most of these correlations range from the mid-.60's to the mid-.80's, probably falling close to the reliability coefficients of individual ratings. The rater reliabilities are not given in the report, although they were evidently calculated since they were used to make certain statistical corrections in later analyses. However, in the light of general knowledge about rater reliability, I judge that most of the scale intercorrelations are within the range of these reliabilities. Further evidence of halo effect is to be found in the correlations of individual scales with the Overall Rating, which are as high or higher.

On the other hand, the ratings yielded low correlations with the other two types of criterion measures, namely, the Job Knowledge Test and the Work Sample. Many of these correlations were too low to reach statistical significance. Moreover, the correlations between ratings and Work Sample scores were consistently lower than those between ratings and the Job Knowledge Test. Yet the Work Sample is more nearly representative of actual job performance and its

scoring requires some use of rating procedures, albeit of a more objective nature. Finally, in the Cartographic Technician sample, in which all three types of criterion measures were obtained, the correlations between Job Knowledge Test and Work Sample were sizeable, ranging from .47 to .55 in the three ethnic groups, while the correlations of ratings with Job Knowledge Test and with Work Sample were consistently lower, ranging from .28 to .42 and from .14 to .37, respectively. A related finding is that the three ethnic groups showed little mean difference in ratings, in contrast to several sizeable mean differences in the two more objective criterion measures.

It is especially noteworthy that these unsatisfactory results were obtained with the rating criterion despite the technical excellence of the construction and use of the rating scales. The traits to be rated were selected and defined after a thorough and comprehensive job analysis. The detailed instructions and the administration of the rating scales to groups of supervisors by project personnel should have reduced common misunderstandings and misuses of the scales. The utilization of specific instances of behavior to anchor each scale, as well as the requirement that all individuals be rated on one scale at a time, are generally recognized procedures for reducing halo effect.

From another angle, special analyses of the ratings were conducted to check on any systematic differences associated with ethnic categories. The results showed several significant interactions between race of rater and race of ratee. For one thing, raters tended to assign higher ratings to members of their own ethnic group. When such ratings were checked against objective measures, however, as could be done with the Job Knowledge ratings and the scores on the Job Knowledge Test, certain group differences emerged. For example, Black raters gave higher mean Job Knowledge ratings to Black than to Caucasian technicians, although on the Job Knowledge Test the Black technicians

obtained lower mean scores than did the Caucasian technicians. With Caucasian raters, on the other hand, the ratings assigned to Blacks averaged slightly lower than those assigned to Caucasians, while the test scores showed a larger difference in the same direction. Thus, in the first case, the differences in ratings and in test scores were in the opposite direction; in the second case, they were in the same direction and the ratings tended to underestimate the test score difference.

Another interaction between ethnic category of rater and ratee appeared in the correlations between Job Knowledge ratings and Job Knowledge Test scores. For example, in the case of Medical Technicians rated by Black raters, the correlation was .50 for Black ratees, but only .09 for Caucasian ratees. With Caucasian raters, the corresponding correlations were more uniform, being .56 for Black ratees and .39 for Caucasian ratees. In general, the Caucasian raters did not exhibit as much variation in either mean ratings or correlations in relation to ethnic category of ratees, as did the Black raters. Still other ethnic differences among ratings can be found in the pattern of correlations between the Learning Ability ratings and the individual predictors. These pattern differences suggest that different aspects of job performance may influence the ratings, and that these differences depend upon the ethnic category of both raters and ratees.

The evidence of these various biasing effects in ratings across ethnic categories helps to explain the relatively unsatisfactory performance of ratings in the previously cited data. The results certainly suggest that ratings are a questionable type of criterion measure for test validation when different ethnic groups are involved.

Apart from these general implications, let me mention briefly some further information I should have liked to see regarding the ratings obtained in the

present study. It would have been helpful to have rater reliabilities, separately reported for traits and for ethnic category of raters and ratees, if possible. I am also curious about the inclusion of several personality traits in the rating scales, since they seem not to have been used in any of the analyses. As for the other traits, I am wondering why predictors were not correlated with such scales as technique, organization, and communication for Medical Technicians; accuracy and dexterity for Cartographic Technicians; and organization, communication, and judgment for Inventory Management Specialists. (Ed. note: Correlations were computed between all predictors and criterion measures, and will appear in the Appendix of the Technical Report.)

I do not agree with the stated justification for singling out the Learning Ability ratings for correlations with all the predictors. "Ability to learn" does not seem to me most closely related to the purpose of most so-called "aptitude tests." On the contrary, the predictors chosen (as well as other aptitude tests) measure what the individual has already learned in some quite dissimilar areas, such as arithmetic computation, vocabulary, spatial visualization, or finger dexterity. It is well established that ability to learn is not a general factor. And ratings on Learning Ability seem a particularly surprising criterion to use when validating tests selected from the Kit of Reference Tests for Cognitive Factors! To be sure, the Learning Ability rating scale may have been chosen for a different and very good reason, such as high rater reliability. What I am questioning is the rationale given to support its choice.

Finally, it could be argued that for the analyses of rater bias, Overall ratings would have been more appropriate than Learning Ability ratings. Overall ratings are the type most commonly employed in industrial validation studies. Moreover, subjective and biasing tendencies are more likely to be

manifested in Overall ratings, in which instructions to raters are the least structured. By choosing a less subjective rating scale, such as that for Learning Ability, the investigators may actually be minimizing the biasing effects they are trying to investigate.

Multiple uses of job analysis

The fourth and last question I should like to raise concerns some special applications of job analysis. In the present study, the results of job analyses served both as a guide in the selection of relevant predictors and as an aid in the development of all three types of criterion measures. These are well established, standard applications of job analysis. I should now like to propose two further applications that seem particularly appropriate in the context of the present conference.

First, I would urge that job analyses be repeated periodically. For a variety of reasons, the functions performed--and therefore the abilities required--in any given job are likely to change over time. To ensure that outmoded requirements are not perpetuated and to keep selection instruments relevant to the job, the periodic reanalysis of job processes appears to be an objective and realistic procedure.

The second application is suggested by a sober consideration of the scope of the present study. I can think of few, if any, real-life situations providing the time, facilities, and technical personnel to permit the kind of test validation represented by this study. Even with the unusual opportunities available in this study, certain planned procedures had to be discarded because of practical obstacles, and some of the subgroups were smaller than desired. In a more nearly typical personnel situation, what, then, can be done to ensure that selection tests are truly valid, or relevant to the job? For this purpose, too, I would turn to a thorough, professional job analysis, followed by a study

of the published research findings regarding the validity of different tests against specific job functions. I would urge that more effort be expended on basic research regarding the specific aspects of behavior measured by different instruments and less on inadequate and inconclusive local validation studies against global criteria of job performance. To me this is perhaps the major implication of both the procedures and findings of the present study.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR EMPLOYERS IN GOVERNMENT

Raymond Jacobson

Director, Bureau of Policies and Standards

U. S. Civil Service Commission

My purpose here today will be to share with you my perceptions as to some of the implications of the studies as I see them for all employers in government. My remarks will reflect the general viewpoints of a public personnel manager and will not refer specifically to the U. S. Civil Service Commission. Since the inception of this project, the Commission's interest in the results has deepened, as we now have responsibility for coordinating all Federal technical assistance in personnel administration to State and local governments. The research, as I see it, is some of the most important which has been done in a practical personnel setting in recent years, and it has served a great need for more definitive information about the impact of employment tests upon various groups in our society. I cannot help but be greatly impressed with the scope of the effort and the diligence with which it was carried out.

I sincerely wish to express the Commission's deep appreciation to the Ford Foundation for its funding and forward-moving efforts to find scientific bases for solving many of our major social problems. I also feel that accolades are due the Educational Testing Service staff for its diligent carrying-out of the research despite a number of administrative and technical difficulties. I would also like to thank the many Federal employees who served in various capacities throughout the course of the conduct of this study. Appreciation is due the agency managers who cooperated in the job analyses and the development of criteria and special tests, and gave of their employees' time to participate in the studies; to the more than 1,400 employees who cooperated in taking the

tests; and to the management and psychological staff of the Civil Service Commission for its efforts in all phases of the study.

As a manager who does not have extensive training in measurement, I must rely upon the psychometricians for detailed interpretations of the study. I am confident, from many years of involvement with personnel measurement psychologists, that they, in analyzing these studies, will have a range of points of view about them, their meanings, and implications. I look forward to hearing these points of view today, and in the months and years ahead. However, as I see the studies in a broad perspective, they each began with careful and extensive analyses of the work being done, and with subsequent psychological inferences regarding the qualifications necessary to do the work. These important steps seem either to have been neglected or not to have been as significant a part of many previous studies of test fairness in other settings. At this point, I should call your attention to the fact that the ETS approach of job analyses employed in these studies as the foundation for qualifications measurement is one which has existed in the Federal structure for many years. I attribute much of our success in minority employment and promotion to this objective cornerstone of the merit system.

As I understand the results, most of the tests chosen on the basis of systematic job analysis were found to be valid for different subgroups, and it was subsequently possible to study the various issues surrounding test fairness. Had the tests not been found to be job-related, it would have been difficult to answer the most important questions about possible test fairness.

Therefore, I see as one major implication for government employers a renewed emphasis on sound job analysis as a cornerstone of fair employment examining. Certainly, public employers should be encouraged to think of job analysis, not just in terms of job evaluation for pay purposes, but more in

terms of developing sound and fair employment procedures. The foundation of merit systems has been that if a job requirement is soundly derived, that is, if it is necessary for effective performance and differentiates among workers in terms of their effectiveness, it is fair. This research seems to me to reinforce the legitimacy of that assumption. It appears to me that more employers, both public and private, should be encouraged to do research on even better job analysis methods and the translation of job analysis data into employment procedures through even more scientific means.

Another implication I see for government employers is related to the difficulty and expense we have seen faced in differential studies such as these. I am both impressed and appalled that it took six years and such a vast amount of money to study these occupations. This is not a critical comment; rather, it is my judgment that most public employers in the country will not be able to follow the path of doing criterion-related validation, particularly differential validation, for various subgroups. That there is a serious money crisis in all governments is not news. Many programs are competing for a limited number of dollars. In this competition, viewing the results of these studies, I question whether your taxpayer dollars would be wisely spent in doing more of these kinds of elaborate differential statistical studies to continue to demonstrate fairness. Please note that I am not recommending the cut-back or abandonment of psychometric research. Quite the contrary. But I am now concerned that it is time to adopt a cost effectiveness approach to the problem of test fairness.

For some time we have been considering the issuance of instructions codifying for the Federal government's employment system the best professional practice in the development of qualifications standards, tests, and other applicant appraisal procedures, and examining methods to assure sound selection and placement without discrimination because of race, color, religion, sex, or

national origin. These instructions have been developed. They place primary emphasis on and demand systematic job analyses as the basis for our qualifications examining practices.[1] The ETS studies confirm that the approach of sound job analyses can be fairly used to build job-relatedness or validity into the selection system.

I am well aware that psychologists have been greatly concerned about the criterion problem for years. These studies highlight the importance of getting good measures of job performance. They also attest to the great difficulty in doing so. Many public employers would find the development of sophisticated work samples, such as those developed for and used in these studies, prohibitive in terms of cost, time, and professional resources.

Another problem is that there appears to be an incipient and growing resistance on the part of both majority and minority group members to participation in such studies. That this occurred in these studies is frankly a surprise to me. It suggests that researchers may be about to run out of time and good will in conducting studies aimed at uncovering group differences. It will require careful thought on my part as to whether to recommend that we in the Federal government ought to risk exacerbating inter-group problems by extending these kinds of studies.

A final implication I see for public employment relates to the role of tests in the whole employment system. I believe these studies have shown clearly that public employers should not resort to flight from well-selected employment tests, nor should they resort to differential use of test results for minority groups. But the studies, although very large in scope, have provided only a small part of the guidance that a personnel manager, devoted to the concept of fair employment, needs. For example, what about the alternatives

[1] These instructions were published in the Federal Register, June 30, 1972.

to employment tests? Are these any more or less fair than tests? The criticism of employment tests has led to considerable research on the fairness of written tests. But we do not know nearly all we need to know about the other aspects of the employment system which impinge upon employment opportunity. For example, we need more solid research on recruiting. Modified recruiting practices for the Washington, D. C., Police Department resulted in an increase of blacks of 228%, while whites were increasing by 47%. This occurred without a change in the written test.

We need more work on the development of practical ways to measure performance. As these research studies have pointed out, the performance appraisal picture through ratings is very complicated. My own view is that the way ratings are done in practice is often worse. Other alternatives to work measurement should command an important segment of our resources.

We, over the years, in the Federal Civil Service, have tended to use written tests less and less in employment decisions. For example, in the Federal structure, about 50% of the initial placements involve non-test methods exclusively. Tests are rarely used in promotion decisions, which account for many more times as many personnel actions as initial hiring. The ETS studies strongly suggest that we may wish to consider reversing the trend away from objective testing.

In summary, I see these as major implications. Sound job analysis made these studies possible. It is apparent that sound job analysis is of utmost importance in providing fair employment. We need to do more to foster professionally developed job analysis systems in public service, and integrate these systems with our employee selection systems. Second, it is clear that differential criterion-related validation should not be accelerated in public service. Such studies are meaningful only if they can be conducted properly. To do so

places an unnecessary, unwieldy burden on many public employers, particularly those in smaller jurisdictions. Finally, validation of tests is not an answer in and of itself to the problems in achieving fair employment. We must look at other aspects of the decision-making process. Only by looking carefully at the whole employment system, evaluating it carefully, and taking steps to improve it can we hope to achieve fair employment. We must allocate our limited resources carefully lest we allow them to be drained in an area such as testing when other barriers to fair employment go untouched. It is a fact that the most significant progress in equal opportunity has been made in personnel systems where personnel decisions are based upon merit principles and objectivity. Nevertheless, public employers must make <u>total</u> plans for achieving fair employment, and the <u>whole</u> personnel system must be studied and improved so we can attain full equal employment opportunity for all Americans.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR EMPLOYERS IN INDUSTRY

Lewis E. Albright
Director, Manpower Planning and Development
Kaiser Aluminum & Chemical Corporation

I am pleased to be invited, as a representative of industry, to comment on these important studies. As many of you know, American industry has made extensive use of paper-and-pencil tests for at least half a century. When used properly in conjunction with other appropriate personnel evaluation procedures, tests have made a significant contribution to improved selection and placement. Unfortunately, industry's record of test usage has not been viewed with unmixed favorability. We have been accused, for example, of using tests to perpetuate conformity and an organization man stereotype (Whyte, 1957). Similarly, our uses of tests have been criticized by others as unwarranted invasions of privacy (Gross, 1962). Still others have charged that the multiple choice format penalizes the most creative individuals because of their ability to see unusual (correct) relationships of supposedly wrong responses with the questions (Hoffmann, 1962).

More recently, industry's use of tests has been the subject of numerous complaints and challenges from minority group individuals who have alleged that unfair tests kept them from obtaining jobs to which they felt they were entitled. Starting with the Motorola Case in the early 1960's, through the Supreme Court's decision in Griggs vs. Duke Power in 1971, and continuing today, these cases will play an important part in determining how tests may be used in the future by all employers, not just those in the private sector.

One problem which has plagued virtually all of these cases to date is the lack of a comprehensive body of knowledge on test validity and test fairness

for minority groups. Much of the evidence has been scattered, often based on small samples and questionable criteria or methodology, and has suffered from the suspicion of reflecting the biases of its originator.

These studies by ETS are particularly welcome, therefore, since they do much to fill this void. They are based on sufficient sample sizes to be reliable. They employ sound methodology in criterion development. The analyses appear to be very complete and well done. I found the cross-ethnic cross-validation technique particularly interesting and one I had not seen before. Even though the studies were done in government installations, the three jobs involved--medical technician, cartographer, and inventory manager--seem similar enough to jobs in industry, such as laboratory assistant, draftsman, and warehouseman, to strike a responsive chord in readers from the business world.

What, then, do these studies tell us which will be of primary interest to industrial users of tests? I think they tell us a number of things and they also raise some questions for all test users.

First, they support the feasibility of multi-location validity studies. Many of us have worried that differences among geographical locations, in terms of differing population characteristics, or variations of job content and criterion measures, might obscure validation results. The relatively high and consistent validity coefficients obtained in the ETS studies indicate that regional differences may not be such a problem. It would be reassuring to see more data, however, on the composition of the samples in these studies, particularly on such demographic characteristics as age, education, and length of service. Any significant regional differences on these variables should be described and explained, of course. Similarly, with regard to representativeness of the samples, more discussion might be devoted to the reasons for more than 100 Inventory Management Specialists declining to participate in the study.

Did their refusal change the composition of the sample in any important ways?

Secondly, I believe all industrial test users will be encouraged to see that careful job analysis and criterion development can pay off in such high validity coefficients. We have been telling ourselves for a long time to give more attention to the "criterion problem," and it is rewarding to find that doing so really makes a difference.

Third, while most of us are accustomed to finding significant differences between predictor means for minorities and Caucasians, we are quite surprised to see similar mean differences, both in direction and magnitude, on the criteria. This finding (and what to do about it) is certainly one of the most important in the entire study. We could dismiss the differences in supervisory ratings as being due, at least in part, to racial bias. But it might be premature to do so without knowing more about the situation. Did the researchers happen to conduct the study at a time of some national crisis, such as the assassination of Martin Luther King, Jr., which might tend to foment distrust and divisiveness along racial lines? Are the biases toward favoring one's own race exhibited in other aspects of the reward structure, including the promotional system and the salary administration program? Are there other evidences of racial conflict in the work setting? In any case, the rating problem does not seem likely to be ameliorated by the usual admonition to "train the raters." The point is that, without knowing more about the situation, it is difficult to suggest solutions. One thing is clear, however: these ratings would almost certainly be unacceptable to the OFCC or EEOC as criteria in a validation study because of the racial bias they now appear to reflect.

Racial differences on the job knowledge test criterion are probably to be expected. The same factors which act to depress performance of minorities on aptitude tests are likely to be at work in the job knowledge tests. Perhaps

for this reason, industrial psychologists have not made extensive use of written tests, and I doubt that they will do so. Other criteria, such as turn-over indices, salary or promotional progress, and productivity data have greater appeal for most of us because they come nearer than a test to reflecting "real life" decisions and actions in most organizations.

The significant differences by race found on the work sample problems seem most disturbing of all the results in these studies because they cannot be explained away as due to method factors or subjective biases. (We cannot overlook the possibility, however, that the Blacks may have less formal education, job training, or work experience than the Caucasians--comparative data on these variables should, I repeat, be included in the report.) These differences imply that Blacks should not be hired for these jobs unless an employer is willing to invest substantial additional training in this group in an attempt to bring their performance up to that of Caucasians and Mexican-Americans. Many private employers might be unable or unwilling to bear these additional costs.

Finally, I believe industrial employers will be most heartened by the ETS data concerning test fairness. There have been previous indications that, in some instances, tests may actually overpredict criterion performance for minorities, e.g., Tenopyr (1967). The present studies provide considerable verification for this earlier evidence by showing rather conclusively that, for these three occupational groups, the regression equations developed on Caucasians were about equally valid for both the Blacks and the Mexican-Americans. This finding, together with the general absence of differential validity in these studies, should do much to blunt the current outcry against testing by those who would interpret any differences in mean test scores as prima facie evidence of unfair discrimination. For this contribution alone, these studies should serve as a landmark for many years to come.

References

Gross, M. L. The Brain Watchers. New York: Random House, 1962.

Hoffmann, B. The Tyranny of Testing. New York: Crowell-Collier, 1962.

Tenopyr, M. L. Race and socioeconomic status as moderators in predicting machine-shop training success. Paper presented at the meeting of the American Psychological Association, Washington, D. C., September, 1967.

Whyte, W. H., Jr. The Organization Man. New York: Doubleday, 1957.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR BLACKS

Roscoe C. Brown, Jr.
Institute of Afro-American Affairs
New York University

Any consideration of the implications of a report for Blacks should begin with an awareness of the hostile social climate within which Blacks and other non-white minorities live in the United States. I am particularly concerned lest these findings which show that the validity coefficients and regression equations appear to be similar for Blacks, Whites, and Chicanos be interpreted to say that tests yield the same results for Blacks and Whites. Since the ETS study deals with prediction and not actual scores, care must be used in interpreting the results of the study so that they are not used to blanket in tests which (though they might yield the same validity [prediction] coefficients for minority groups) do yield different scores. I believe that ETS also has a responsibility to emphasize this caveat. This study does not vindicate tests as a "color-blind" technique; the study merely says that with our present state of knowledge we do not find any measurable differences in prediction. But we should recall that even though there are no differences in prediction, there are differences in actual raw score test results. We must continue to attempt to account for these differences if we are, in fact, to say that tests are color-blind.

My comments on the technical aspects of the report will be brief because they have been covered by other speakers, and also because the findings were not particularly unexpected by me. The fact that there is a difference between Blacks, Chicanos, and Caucasians on the various aptitude measures does not surprise me. This observation is based on several years of experience, during

which I have studied the role of intellectual and non-intellectual variables in the prediction of school achievement of both Black and white populations. Likewise, the relative similarity of the regression equations for the Black, Chicano, and Caucasian groups is not surprising to me. While some of the more recent commentators on the misuses of tests with minority populations have suggested that there should be different predictive equations for minority groups, a perusal of past literature indicates that there is really no reason to believe this. Assuming that the predictive variables and the criterion variables are reasonably reliable, the only expected difference might be the one that was shown in this study, namely the regression lines for the minority groups tend to be at a lower level than the regression lines for the majority groups. The reason for this, in my opinion, is the accumulative effect of variables which result from the societal context, variables which are not usually measured and possibly, at the present time, cannot be measured in a reliable fashion. I use the term "incremental bump" to describe the additive effect of these variables. An example of the type of variable that causes an "incremental bump" is the inter-personal interaction which is required in solving various problems on the job. Frequently, when Blacks and other minorities, who have not had peer relationships with whites, are faced with face-to-face and eyeball-to-eyeball confrontations about problems which have a cognitive basis, they tend to be less aggressive, less competitive, and less innovative in searching out various solutions to practical on-the-job problems than their white counterparts. This causes them to be slower in developing the type of refined behavior in a particular job that would lead to a higher overall rating. I think that this is one of the external factors that causes the "incremental bumps" which leads to higher performance of white populations on both prediction and criterion variables.

A major problem in the study (this is certainly not the fault of the design of the study, but rather the nature of the situation in which the study was conducted) is the lack of control for the degree and quality of on-the-job training experienced by the participants in the study. It really is unreasonable to hire people who have differential performance or aptitude scores at the time of recruitment, put them on the job, give them practically no training, and then expect them to perform at identically the same level of people who came in with somewhat higher levels of aptitude. This is, in fact, what we have in the present study. Although it might be suggested that there is some organized learning that takes place on a job from day to day, month to month, and year to year, the fact is that unless there is a very specific program that focuses on the particular weaknesses or particular job problems of particular individuals, the probabilities are that workers with low entry scores will perform at a minimum level. In a sense, the selection of the sample, which reflects the pools from which the researchers had to draw, is biased on socioeconomic and educational factors. The best example of this is the fact that larger numbers of white medical technicians in the sample had scientific training in college, while the Black sample of medical technicians contains a larger number of people who majored in the social sciences. While there is nothing esoteric about scientific training for a scientific career, there are certain little skills that one gains through formal scientific training which might contribute to better job performance. Since the Black population and the white population do not start from the same point, you have the basis for differential job performance--a difference which must be overcome with training as well as experience.

I think the main implication of the study for Blacks and other minorities is that we must look for another concept in terms of predicting and evaluating

minorities for positions which have specific job performance requirements. Since none of the predictors was particularly effective in identifying those minority people who would perform at levels greater than you might normally expect from their aptitude scores, I suggest that the basic approach to use in selecting minorities is one which involves establishing a relatively low cut-off point for entry level selections, followed by on-the-job training. (Incidentally, I want to compliment the study group for developing and extending the concept of using aptitude measures that are job-related. In some instances, I think they had to stretch a point in order to select aptitude measures that were job-related, but nonetheless they should be congratulated for their efforts in this direction.) Ideally, if one could identify two or three simple or reliable tests which reflect at least the very basic skills necessary for a job, it should be possible to use the approach of selecting from a pool of minimally qualified people and then begin at the time of placement to conduct an on-the-job training program, both for performance of the job at the particular level at which the person is being employed and for promotion and upgrading. One of the complaints of minorities is that in order to beat the ethnic numbers game, some organizations hire large numbers of minorities at entry level jobs, do very little to upgrade them, and then give the excuse that the minorities just don't have the skills to be promoted. I maintain that, within certain broad outlines, people who are selected using criteria that have some relevance to the job do, in fact, have a potential for higher positions which can be developed through training. Since society has provided neither adequate education nor social support for programs to improve the skills of minorities, I suggest that organizations like the Civil Service Commission and the larger corporations which are under affirmative action plans should adopt a model of selection and upgrading that includes training as one of its most

important elements. A final part of the model which I am suggesting here (one that will take several years to implement) is that after X number of individuals have gone through the selection and training process and have performed on the job at various levels, we attempt to identify certain criteria or characteristics which are associated with the quality of their performance at that time. The assumption here is that the minority group individuals will perform, after adequate training, at levels which are consistent with the cross-sectional white population. If this be the case, and I personally believe it will be the case, we can then look at the personal and performance characteristics they have at that time to see what new relationships might be found between these characteristics and job performance. Until such time as the educational and social opportunities are made equal for minority and white populations, we should use the approach of selecting personnel using a relatively low score on entry level criteria and then training them to the level of skill required by the job.

Another problem with the study is that while it deals very effectively with cognitive measures and job performance measures, it does not deal with what might be the most important factors in job performance and their relationship to supervisors' ratings and upgrading for promotion, namely, non-cognitive factors such as persistence, the ability to get along with one's colleagues, volunteering, spending a little extra time to do a job well, correcting errors without rancor and hostility. These are the things that tend to be involved in getting ahead in any area of business. It might be suggested that these factors are not as significant in some of the technical types of Civil Service jobs. I am inclined to question the allegation that these factors do not apply to Civil Service jobs as well, because just as differences in ratings of supervisors based on the race of the supervisor were found in the study under discussion, differences in the evaluation of performance in even technical areas are

functions of certain non-cognitive characteristics. Therefore, it is imperative that studies of job performance of minorities include some non-cognitive factors. Admittedly, these are more difficult to measure reliably and validly but, in my opinion, we cannot continue to ignore them if we are going to approach the complex problem of predicting the job performance of minorities in a way that reflects the entire picture.

I believe that this conference and this very significant study should lead to significant modifications of some superficial views about prediction of minority group job performance. Some people have suggested that there ought to be some magic formula or magic equation that can be used to identify capable minorities from the larger pool of minority applicants. Unfortunately, as this study shows, there is no magic formula. It is unreasonable to expect that some "magic bullet" will come along to solve the problem. We should stop using testing as the basis for self-fulfilling prophecies. Namely, when people are selected with low test scores, and then perform at commensurately low levels, someone says, "I told you so." Even though the level of performance of minorities compared to majorities on the tasks in this study tends to be lower, the average performance on the part of both groups is still quite competent. Job performance is a reflection, in part, of the effect of the external environment which the minority group workers experienced prior to even being hired. I believe that the model that I suggest, a model which emphasizes training, should obviate the self-fulfilling nature of the prophecies where tests are used to select and to predict low levels of performance.

An interesting part of my experience is as the Chairman of the Examining Board of the Manhattan and Bronx Surface Transit Operating Authority (MABSTOA), a quasi-official agency in New York City which operates the buses in Manhattan and the Bronx under the general supervision of the New York City Transit

Authority. Since MABSTOA is a quasi-official agency (due to the fact that the bus lines involved were taken over from private ownership and the final disposition of their status is still in limbo), an examining board of citizens was established to monitor and oversee the selection and promotion practices in this agency. My colleagues on the Board (William Mulligan, the former Dean of Fordham University Law School, now a Federal judge; Professor Sidney Mailick of the N.Y.U. Graduate School of Public Administration; and Dean William Moore of Fordham University Law School) and I have been able to utilize some of the principles mentioned earlier in this paper. We have stimulated the use of tests which have some element of job relationship for selection for entry level jobs and then have worked with MABSTOA to increase the amount and quality of in-service training. A major problem we face is that it is very difficult for an operating agency to provide on-the-job training for the job itself and for upgrading and, at the same time, perform its operational role which, in this case, is to have an adequate number of buses on the streets running on schedule. A considerable number (over 50% in some categories) of the persons selected for both entry level and promotional positions using job-related problems have been minority group people. Evaluations of their supervisors show that performance of personnel selected in various job categories has been considerably better than it was before the Examining Board. The Examining Board can't take complete credit for this because when the bus lines were under private management their recruitment and promotional programs were largely based on informal arrangements and personal contact, a fact that suggests the pool from which private management was drawing was not of the same quality as the pool from which the Board is drawing, now that there is a public announcement of the selection and promotion process. I only mention this experience to reinforce the point that when an agency

actively seeks to increase the number of minority workers it employs and plans to move them upward in that agency, it should seek out approaches that are attuned to the realities of the agency and attempt to strike some new ground in terms of selection and promotion procedures.

I believe that the ETS study is an important study, not so much because of the actual results of the study, but because of the many issues that it raises. It eliminates certain areas as major areas of concern, but it also suggests many other areas that have not been adequately explored or identified. As a Black who has been involved in the educational research and measurement process for some years, I am not prepared to say, "Down with all tests." On the other hand, I am in no way prepared to accept the ways in which tests have been used in the past, or ways in which it has been suggested that they be used, to select minority workers and to predict their success on the job. Clearly, much more needs to be done in the area of the assessment and prediction of the job performance of minority group members. It cannot be argued with any cogency that no method of selection should be used. Obviously, some method of selection and promotion should be used. The question before us, then, as scientists and social theorists, is how we are going to do this with equity, with reliability, and with some degree of accuracy. This is the real reason why scholars, scientists, and administrators concerned with the public domain come together in gatherings like this--to examine our problems and to determine what we need to be thinking about in order to solve them. As I indicated in my opening remarks, this study has implications for society as a whole, not just for Blacks. I have given my own points of view as a Black person who is competent in the area, but I also have reflected my concern as a scholar who feels that we have attributed validity to various procedures and mechanisms that may not be warranted at their present stage of

development. Again, I want to congratulate the sponsoring agencies and the participants in the study for an in-depth exploration of a very complex phenomenon, an explanation which has opened some vistas for me personally and, hopefully, for a nation that is attempting to deal with the important issue of equality of opportunity in all fields of endeavor.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR SPANISH-AMERICANS

TOWARD MORE "TESTEE CONSTRUCTION THEORY" AND LESS "TEST CONSTRUCTION THEORY"

Edward J. Casavantes

Executive Director

Association of Psychologists for La Raza

I

This critique is in response to a request from the Educational Testing Service (ETS) for a review of selected chapters of their report, "An Investigation of Sources of Bias in the Prediction of Job Performance," for special relevance to the Spanish-speaking community.

As a Chicano psychologist-sociologist, I am not sure that I can speak for Puerto Ricans and for Cubans--and perhaps I do not even speak for many Mexican-Americans. Nevertheless, there are a number of serious concerns that need to be articulated, which I believe are relevant not only to Spanish-speaking minorities, but to minorities in general.

I have adopted the viewpoint that to the degree that I am being asked to review this manuscript from the viewpoint of a social scientist from a minority background, then to that degree I have to be concerned not primarily with the tests but with the people taking these tests. This is the rationale I use when I say we ought to pay more attention to the "construction of the testee"--that is, his make-up: his background, the discrimination he has faced, his language, his culture, his poverty level, his lessened opportunities, his traditions--and less attention, especially in projects like this one, to the "construction of the test."

Essentially, then, my tasks boil down to two somewhat overlapping elements: (a) to look at those things that ETS did not look at in terms of sociologic-environmental factors which may affect the sources of prediction bias, and (b) to take a closer look at the attributes of the incumbents themselves (the "incumbents" are those individuals, both minority and Anglo, who participated in the ETS study).

Although I am familiar with statistics and experimental design, I felt that other members of the panel of reviewers, who are legitimately respected in the specific areas of statistics, experimental design, and test construction theory, would do a far more adequate job of appraising the technical approaches used by ETS in its analysis of the obtained data. My supposition was well-founded, and their critiques of the existing data, as far as I am able to determine, are excellent.

Thus, the present issue--for me, at least--is not whether the statistics or the statisticians are adequate. I, for one, am willing to give the numerical and methodologic processes used in the present study a clean bill of health. But, I don't feel that this is the real problem in a study to determine the adequacy of prediction of (relatively) standard tests with minority people.

The real problem is whether the numbers that were so extremely well gathered and then extremely well manipulated are values that represent--that is, are sufficiently isomorphic with--the real life circumstances with which the incumbents have had to deal. Thus, a score of 62 for an Anglo may simply not mean the same as a score of 62 for a Chicano working side-by-side that Anglo. One of these scores of 62 may have been much more hard-won than the other.

A quotation by the philosopher Suzanne Langer (1942) seems particularly

appropriate at this time; the thirty-year lapse between the time she said it and today accentuates it and makes it all the more appropriate:

> The faith of scientists in the power and truth of mathematics is so implicit that their work has gradually become less and less observation, and more and more calculation. The promiscuous collection and tabulation of data have given way to a process of assigning possible meanings, merely supposed real entities, to mathematical terms, working out the logical results, and then staging certain crucial experiments to check the hypothesis against the actual, empirical results. But the facts which are accepted by virtue of these tests are not actually observed at all. [My emphasis]

The Chicano social scientist views his people first of all as people, ordinary human beings, but human beings with some very unique attributes. Some are negative attributes: high proportion living in poverty settings, having had poor medical attention; low educational level, having felt discrimination; and known lack of opportunity. Others are affirmative: two languages, two cultures, two histories, two life styles. Other attributes of many Chicanos should be irrelevant to success in life, but may, under certain adverse conditions, affect them: their Catholicism, their being darker of skin, their having "a Spanish accent," their preponderance in the five Southwestern states.

Therefore, I must view the ETS study from the above perspective. And, from this perspective, the ETS study data and conclusions are highly suspect. ETS failed to look at many factors which without doubt entered into the production of the numbers ETS later analyzed in a very adequate manner.

Perhaps it is important to note that ETS could have known about many of the factors about which I will later voice concern. Few elements I will mention were not available and potentially "knowable" to ETS at the time of the study. Why ETS did not look carefully at these factors is a serious matter for ETS to consider in any future studies of this nature.

II

About three years before ETS gathered data on Inventory Managers at Kelly Air Force Base, San Antonio, Texas, the U. S. Commission on Civil Rights conducted an investigation into personnel problems of that Base.

> On November 7 and 8, 1967, the Texas Advisory Committee to the U. S. Commission on Civil Rights met in closed session at the El Tropicano Hotel in San Antonio, Texas, to receive information on employment practices and policies at Kelly Air Force Base.
>
> During the two days, the Committee received information from 40 persons, including military and civilian officials of Kelly Air Force Base, a representative of the U. S. Civil Service Commission, local Mexican American leaders, representatives of two trade unions with members at the Air Base, and white-collar and blue-collar employees of the Base. (Kelly AFB TAC Report, p. iv, June, 1968.)

A gross overview of the employment ratios for the three major racial-ethnic groups shows that, interestingly, the (1966) population distribution of San Antonio as a whole did not differ significantly from the distribution of all Kelly AFB employees. As can be seen from Table A, these were:

About 50 percent Anglo
About 44 percent Chicano
About 6 percent Black

Table A

MEXICAN AMERICAN AND NEGRO EMPLOYMENT AT KELLY AIR FORCE BASE

June 30, 1966

Category	Total	Mexican American		Negro	
		No.	Pct.	No.	Pct.
All plans	22,293	9,764	43.8	1,428	6.4
Wage Board	12,346	7,035	57.0	1,080	8.7
Class. Act	9,929	2,729	27.5	348	3.5

(Adapted from Table I, p. 2, of the Kelly AFB TAC Report.)

As can also be seen from the breakdown by "Wage Board" (mostly "blue-collar" occupations) and the Classification Act (mostly "white-collar"), Mexican American and Negro employees predominate in the blue-collar group. These figures instantly raise questions about fairness in employment practices.

A more detailed breakdown of the job situation at Kelly AFB in 1966 is given in Table B. It is clear that, beginning with GS-11 grade (professional) with 11.6 percent Mexican Americans and 0.5 percent Black, and steadily declining until they literally disappear above Grade GS-14, the situation for minorities at Kelly AFB was, in 1966, very questionable.

When grade or salary is considered for both white-collar and blue-collar workers, and using the 44 percent Mexican American and 6 percent Black San Antonio population as a base, the disadvantaged position of minority workers stood in sharp contrast to that of Anglo employees. Among Mexican American white-collar employees, 69 percent (157 percent of parity) were in the lowest grades GS-1 to GS-5, for which the initial annual salaries (in 1966) were $3,609 to $5,331. Of the Negro white-collar workers, 71 percent were in these GS 1-5 grades. The higher the grade, the fewer the minority group workers, whether in white-collar or blue-collar jobs (Kelly AFB TAC Report, p. 2).

In the blue-collar occupations also, the better paying jobs were steadily fewer for minorities as annual salary increases in June 1966, as shown in Table B. Again using 44 percent Chicanos and 6 percent Blacks as a base, it is clear that these minority employees have been very much discriminated against in advancement opportunities. Only 2.4 percent Blacks and 32.4 percent Chicanos were earning as much as $7,999 per year. Only <u>one</u> Black (0.6 percent) earned as much as $8,999. No Chicanos, out of a total Kelly AFB Wage Board Chicano force of 169, made over $11,999.

Table B

MEXICAN AMERICAN AND NEGRO EMPLOYMENT AT KELLY AIR FORCE BASE IN UPPER GRADE AND SALARY LEVELS

June 30, 1966

CATEGORY	Total	Mex. Amer.		Negro	
		No.	Pct.	No.	Pct.
Class. Act					
GS-11	1,220	142	11.6	7	0.5
GS-12	657	34	5.1	4	0.6
GS-13	216	8	3.7	1	0.4
GS-14	57	1	1.7	1	1.7
GS-15	18	0		0	
GS-16	1	0		0	
Total	2,169	185	8.5	13	0.6
Wage Board					
$ 7,000- 7,999	416	135	32.4	10	2.4
$ 8,000- 8,999	155	27	17.4	1	0.6
$ 9,000- 9,999	73	5	6.8	0	
$10,000-11,999	24	2	8.3	0	
$12,000-13,999	5	0		0	
$14,000-15,999	1	0		0	
Total	674	169	25.1	11	1.6

(Adapted from Table II, p. 4, of the Kelly AFB TAC Report.)

For these reasons, we have to take a very serious look at the so-called "incumbents." We know they do not represent "average" Chicanos; further, the "average" Chicano is lower than the "average" Anglo. The same is true for Blacks. Thus, on this basis alone, the equating of test scores and the rating of job incumbents by supervisors is highly suspect.

The Kelly AFB TAC Report attempted to address itself to some of these issues by calling attention to "Problem Areas." These are, in part (pp. 5-8):

> During the 2-day session, the Texas Advisory Committee heard numerous statements concerning employment practices and procedures at Kelly Air Force Base which minority group persons felt were discriminatory, or worked to their disadvantage. Some Federal officials appeared to believe that the inequities for the most part were due to educational-cultural differentials between the minority and majority populations. Many community leaders, however, disputed this view and urged the [USCCR Texas] Advisory Committee to investigate and carefully consider each problem area.
>
> The major complaints concerned promotion to supervisory positions and the higher pay grades and levels. Complainants alleged that personnel policies and practices, combined with individual prejudices and preferences, resulted in "a system" which made it difficult for the minority worker to be promoted. Similar concerns were expressed by a Mexican American consultant who had reviewed the Equal Employment Opportunity (EEO) Program at the Base. [My emphasis]
>
> There were specific complaints about inequities within the following factors relating to promotion procedures:
>
> 1. The Learning Ability Test. This test was one of three major factors which determined whether a worker gets on a profile, or list, of those eligible for promotion. The other two determinants are: experience and training, and the supervisor's appraisal.
>
>   Complainants stated that the Learning Ability Test, a standard Air Force test, reflected a strong middle-class bias and was unfair to minority groups. They also alleged that the test had no relevance to job performance...
>
>   Management officials at Kelly Air Force Base acknowledged difficulties with the test and reported that it had been discontinued for most unskilled positions...

2. The Supervisor's Appraisal. Mexican American citizens complained to the Committee that many supervisors are prejudiced against minority groups. Others alleged that the supervisor's appraisal is a very objective rating, and minority group identification often was given greater consideration than actual job performance...[This allegation is given substance by ETS's own findings. ETS found each racial-ethnic group gave itself higher ratings. To the degree that there are more Anglo supervisors--and there are--more Anglos will receive favorable ratings.]

3. The Pass-Over. Closely related to the above, and in effect another aspect of the supervisor's appraisal, is the pass-over. Complainants alleged that many Mexican Americans and Negroes who were able to get on a profile and thus be included in the "area of consideration" (the top five names) were passed over by supervisors in the final selection process. Hence, it was complained that "supervisors get two cracks at us"--the first time in getting on the profile and into the area of consideration, and again in the final selection of the person to be promoted. Among those in the area of consideration, the supervisor's decision solely determines who gets promoted. Several Mexican Americans stated that they had been on profiles for considerable periods of time and had been passed over for majority group employees.

4. Promotion Evaluation Pattern (PEP). The PEP is a statement of the requirements for a position which is developed at the Base when Civil Service Commission requirements are too broad to cover a particular job. The PEP is usually written by the Personnel Office in conjunction with the supervisor. Mexican Americans complained to the Committee that in some instances a biased supervisor, with the assistance of the Personnel Office (where few minority workers are employed), could "tailor-make" the PEP to insure the selection of a particular individual as the most qualified.

Among the Commission's Advisory Committee's findings were:

> The Committee finds, and the statistics in this and other Government reports substantiate, that there are broad and glaring inequities in the distribution of supervisory and higher grade positions among Mexican Americans and Negroes, and white citizens of non-Mexican background. [My emphasis.]
>
> The continued existence of these inequities, whatever their original source and the current explanations, constitutes a major and pressing problem for a large number of Kelly Air Force Base employees and, indeed, to Mexican American citizens and leaders in the San Antonio community and throughout Texas.

Two reminders need to be made in order to "set the stage" for the subsequent evaluation of the study incumbents: first, that the possibilities for advancement are very different for Anglos and for minorities at Kelly AFB; and second, that Kelly AFB is not being "picked on," since, as will be documented later on, the Veterans Administration also has been uneven in its treatment of minorities. Kelly AFB was used as an example only because, fortunately, minority employment data and testimony were available for it.[1]

We who work in the area of minority relations and civil rights have found that these inequities exist almost everywhere, and that all one has to do is scratch the surface to find these inequities. Or, to put it another way, "test construction theory critiques" have not often incorporated these types of interpretive data, either because of unwillingness to get into troublesome matters of ethnicity and race, or because data on racial-ethnic problems were not available for the population being used in the study. Fortunately, in this case, both factors are present.

We turn now to some possible specific interfaces between the Civil Rights Texas Advisory Committee and ETS study findings. ETS reports:

> A decision was made to test [for the ETS study] primarily at grade levels 9 and 11, the journeyman levels in inventory management, after progress through the GS-5 and -7 training periods. (Entry into the 2010 classification is at grade 5, with progress to grade 7 and then grade 9 within a prescribed period, subject to satisfactory performance.) A number of inventory managers in GS-7 were included in order to increase the ethnic samples.

Several problems are easily seen. First of all, there were very few GS-11 Chicanos or Blacks. Were these very few being compared with the more abundant Anglos? The phrase "a number of inventory managers in GS-7 were included in order to increase the ethnic [only?] sample" clearly substantiates that there

[1] Editor's note: The Mexican-American Cartographic Technicians included in the study were from the Army Topographic Command at Fort Sam Houston, San Antonio.

were few minority GS-11's and above. In addition, it may be that a disproportionate number of GS-7 minorities may have been compared with higher-ranking Anglos. What these two processes alone will do to the prediction formulas may be enough to invalidate them. Or, more accurately, not to the formulas, for these are probably mathematically accurate, but to the meaning and validity of the prediction formulas, for here the possibility that we are comparing apples with oranges is extremely high.

My suspicions, ironically, are again aroused by the very presence of some high GS-level Chicanos. What special attributes did these Chicanos possess? We don't know, but they must have had much on the ball, or else they would not have made it this high. Or, is it possible these few had become so "Anglo-ized" that they did not represent "average" Mexican Americans? Again, I don't know, and I admit it. ETS doesn't know either, but ETS reports data from these incumbents as if they did not represent a possibly highly unique group.

I noted with great interest that ETS was not able to locate in the Veterans Administration hospitals "enough Mexican Americans" for a Medical Technicians sample. This simple declarative statement literally wipes out what may be a much more important source of test bias than the elements found by ETS; this non-existence of VA Chicano medical technicians is a more vital issue than the sophisticated analytic system ETS attempted.[1] Interestingly, even in Los Angeles, with the largest concentration of Chicanos in the whole country, over a million, there were not enough Chicano medical technicians for analysis, but

[1] Editor's note: In the 30 hospitals across the U. S. where Medical Technicians were tested, there were too few Mexican-Americans to comprise a statistically viable sample. However, they were not "non-existent." Mexican-American Medical Technicians in VA hospital laboratories in the Southwest as of 1967: Tucson, Ariz., 2 of 11; Albuquerque, New. Mex., 3 of 11; San Francisco, Calif., 1 of 15; Dallas, Tex., 0 of 28; San Antonio, Tex., 0 of 4; Los Angeles, Calif., 2 of 32; Long Beach, Calif., 3 of 39; Phoenix, Ariz., 1 of 5; Livermore, Calif., 1 of 5.

there were, evidently, enough Black medical technicians. Surely, this phenomenon should have caused suspicion.

Fortunately, there are data available today to be able to answer this question, at least in part. Only gross numbers are available from the total VA employment structure, but even this large overview may give the reader the type of perspective necessary to understand our concerns.

A recent U. S. Civil Service Commission publication (1970) reveals minority employment data for the Veterans Administration as a whole (see Table C). The most obvious fact for our concern is that Blacks are not "discriminated" against in the (lower) ranks of the VA. Although Blacks represent about 11.1 percent of the nation's population, they represent 26.1 percent of VA employees, a representation which is over two-fold their national representation. This phenomenon, just like underrepresentation of Blacks in certain agencies, should have merited the attention of ETS, for special selection processes were clearly operative here, even if in favor of Blacks.

Nevertheless, the apparent VA "favoritism" for Blacks quickly vanishes as grades rise. Past Grade GS-11, there is an exceedingly rapid decline in proportions, and only between 2 and 3 percent of Blacks hold the higher positions. Once again, we begin to question promotion policies. Who, then, were the Black medical technicians? Did they represent a very special group of Black individuals? Or were they demographically and sociologically essentially equivalent with their Anglo medical technician "peers"?

For the Spanish-speaking, of whom nationwide some two-thirds are Mexican American, the VA picture is bleak. Although the Spanish-speaking constitute some 5 to 6 percent of the national population, they represent only 2.1 of the VA personnel, roughly one-half to one-third population parity (see Table C).

To what degree these gross national figures for the VA are mirrored in

Table C

1969 MINORITY GROUP STUDY

VETERANS ADMINISTRATION

FULL-TIME EMPLOYMENT AS OF NOVEMBER 30, 1969

Pay System	Total Full-Time Employees	Negro		Spanish Surnamed		American Indian		Oriental		All Other Employees	
	Number	Number	Pct.	Number	Pct.	Number	Pct.	Number	Pct.	Number	Pct.
Total All Pay Systems	146,523	38,205	26.1	3,032	2.1	335	.2	1,089	.7	103,862	70.9
Total General Schedule or Similar	113,110	24,719	21.9	2,026	1.8	234	.2	952	.8	85,179	75.3
GS- 1 thru 4	45,402	15,688	34.6	964	2.1	127	.3	125	.3	28,498	62.8
GS- 5 thru 8	31,051	6,584	21.2	523	1.7	55	.2	278	.9	23,611	76.0
GS- 9 thru 11	21,787	2,047	9.4	226	1.0	33	.2	270	1.2	19,211	88.2
GS-12 thru 13	8,257	288	3.5	68	.8	13	.2	76	.9	7,812	94.6
GS-14 thru 15	6,381	107	1.7	245	3.8	6	.1	203	3.2	5,820	91.2
GS-16 thru 18	232	5	2.2	0	0.0	0	0.0	0	0.0	227	97.8

[Adapted from: U. S. Civil Service Commission. "Minority Group Employment in the Federal Government." November 30, 1970. (SM70-708) U. S. Govt. Printing Office. Washington, D. C. (p. 283)]

the ETS sample is, of course, not known, but the availability of Blacks and the non-availability of a sufficient number of Chicano medical technicians for study by ETS certainly square with what one might expect from the national figures.

Again, it is hard for me to say with precision to what degree bias has been introduced by the VA hiring and promotion policies, but bias I am sure exists. I can't account for it; but neither does ETS account for it when it publishes numbers about this obviously biased sample.

Clearly, ETS has its disclaimers, but the disclaimers do not prevent it from publishing the figures as it found them. It is in the publishing of them--in the very act itself--that these figures gain legitimacy.

I can predict right now that it is the ETS figures and conclusions--and not the criticisms of them, such as are being presented in this paper--that are going to be bandied about in academic circles, in Congressional hearings on the validity of tests for minorities, and in educational circles where massive student group testing goes on with only mildly increased concern.

III

This section is somewhat less detailed. It attempts to cover other points which, individually, may not affect test bias significantly. However, in combination, and especially when added to the concerns expressed earlier, their cumulative effect may be very serious indeed. My feeling, then, is that their effect on test bias is additive, and that this is the proper perspective with which to view them.

It appears that the specific occupational categories were selected for either easier availability or for the logistic convenience of ETS. Other considerations stated by ETS are that these choices facilitated experimental

design and statistical analysis. These are among the least worthy criteria for adequate test development in an area specifically designed for original research into the appropriateness of psychological tests on minority peoples. ETS should have elected to seek further.

Relative to the selection of occupations to be studied, the selection of Inventory Management Specialists is appropriate, since the implications from this type of work are transferable to many types of merchandise handling. However, the choice (for convenience?) of jobs such as Cartographic Technician for analysis is almost useless, for this job does not affect 99.99+ percent of Mexican Americans. Thus, in the latter occupation, even if the study results were valid (and clearly we do not feel they are), their utility would be almost non-existent, for transferability is almost non-existent.

Elsewhere in the report, ETS tells us that the test batteries that were selected for use in the study were selected--along with other reasons, it is true--because they were:

a. Short tests. (Doesn't this, in general, lower the reliability and the validity of tests?)

b. Separable into halves. (Presumably to permit easier computation of split-half reliabilities, etc. Also, this makes "short" tests even shorter.)

c. Because they have "known factorial content." (The Lesser, Fifer, & Clark study (1965) and other studies clearly show different patterns of cognitive styles for different ethnic groups, thus, making the "known" factorial content of the tests possibly "not known" for the ethnic populations being studied. ETS does not empirically document that the factorial content of the tests they used were the same for all ethnic groups.)

So, even the selection of the tests used, and from which the data were later analyzed, is suspect.

A fourth problem arises from the fact that the individuals involved in the study were told that its purpose was to study testing and rating procedures of minority peoples. The obtained ratings were then accepted as "true-to-life" by ETS. The United States Commission on Civil Rights, in its Mexican American Education Field Study, in 1971, found a consistent pattern of teachers favoring Black students when one of its staff, a female Black research assistant, was in the classroom making her observations. The pattern of responses by the teachers was so pronounced--clearly a response of "being fair to Black students"--that entire sets of data from that one Black observer had to be discarded. What effect did this "give-ETS-what-it-is-looking-for" phenomenon have on the ETS data is not known, but it is not accounted for in the data presentations by ETS.

The fifth point is important more because of its uniqueness than because it may have significantly affected the numerical data in a highly systematic manner. It is common knowledge that minority people and poor people have a higher arrest and conviction rate than do whites and middle-class people. Thus, with regard to Cartographic Technicians who were minority, and who had to have security clearances to work on classified maps, clearly many of those with "arrest and conviction" records and who may have applied for this job were probably eliminated. In all likelihood, the differences here are small. But that is not the point. The point is that there are undoubtedly a score of other such "minor" factors which may have kept minority peoples not only out of Cartographic Technician slots, but also out of many other positions.

I

Our conclusion, therefore, is that many factors--uneven hiring practices

by Kelly AFB and by the Veterans Administration, inequitable promotion procedures, the selection of logistics that were "convenient" to ETS in the execution of its study, certain security investigations, in one case, the choosing of an almost irrelevant job classification to study, the technical characteristics of the tests themselves--all, in additive fashion, almost without doubt created a situation that was unequal for minority--both Chicano and Black--workers before the study had even begun. It was the attributes of these unevenly-selected and unevenly-placed peoples that ETS then studied in an "even" manner. And, it is these data from this highly questionable set of circumstances that ETS now presents, and, presumably, now hopes we can accept.

It is not important whether the findings of the present ETS study are "positive" or "negative," for either circumstance would be equally suspect. The present ETS findings are unacceptable as scientific evidence that present psychological tests are adequate and/or equivalent measures of prediction of job performance for minority peoples as compared to Anglos.

Our recommendations to ETS are very simple:

a. Hire minority people--a broad spectrum of social scientists, union workers and officials, school officials, students, legislators and other government workers, civil rights workers, and even potential incumbents to be studied--to help in the original design (not just to obtain "permission") for a study such as this one. It is at this stage that their help is most valuable.

b. Do not hire minority people to evaluate what is, for all practical purposes, a fait accompli, for this can but lead to frustration for all parties concerned.

c. Pay far more attention from now on to "testee construction

theory"--that is, to the attributes, the history, and the social circumstances surrounding the individuals who are taking the tests--than to "test construction theory."

References

Langer, S. K. Philosophy in a New Key. Cambridge, Mass.: Harvard University Press, 1942.

Lesser, G. S., Fifer, F., & Clark, H. Mental abilities of children from different social-class and cultural groups. Monographs of the Society for Research in Child Development, 1965, 30 (Whole Issue No. 4).

U. S. Civil Service Commission. Minority Group Employment in the Federal Government. November 30, 1970. (SM70-708) U. S. Gov't. Printing Office, Washington, D. C.

U. S. Commission on Civil Rights. Employment Practices at Kelly Air Force Base, San Antonio, Texas. A Report of the Texas Advisory Committee to the U. S. Commission on Civil Rights. June, 1968.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR GOVERNMENTAL REGULATORY AGENCIES

Robert M. Guion

Professor of Psychology

Bowling Green State University

Federal regulatory agencies exist to implement national policy. Unfortunately, national policy is not always clear. It is determined in part by Congress, in part by the President, in part by the Courts, and in part by the agencies themselves. Laws, Orders, Decisions, and Guidelines are written at different times under different circumstances by different people; it is natural that the results are somewhat ambiguous. Where there is confusion as to policy itself, there will be confusion about the implications for policy of any given set of facts.

With regard to equal employment opportunity, national policy would seem to be fairly straightforward. The Civil Rights Act of 1964 (as amended) says: "It shall be an unlawful employment practice for an employer to fail or refuse to hire...any individual...because of such individual's race, color, religion, sex, or national origin; or to...classify...applicants for employment in any way which would deprive...any individual of employment opportunities...because of race, color, religion, sex, or national origin."

Executive Order 11246 is similar: "The contractor will...state that all qualified applicants will receive consideration without regard to race," among other things. And the Supreme Court said: "Congress has not commanded that the less qualified be preferred over the better qualified simply because of minority origins."

Congress was also fairly precise about what national policy is not. For example, it is not national policy to overlook bona fide occupational

qualifications, or to endanger national security, or to encourage the employment of Communists. Specifically: "Nothing...in this title...require[s] any employer...to grant preferential treatment to any individual or to any group because of the race, color, religion, sex, or national origin of such individual or group on account of an imbalance which may exist..."

From these statements, it seems clear that national policy requires employers to consider each individual on his own merit. That should mean that the employer must find valid means of determining individual merit. That is, he should validly predict (implicitly, at least) how well each individual applicant will do if hired and base decisions on that prediction. If the accuracy of prediction for the individual is enhanced by treating race as a moderator, then race would be properly considered; otherwise, predictions should be made independently of group identifications.

This is an attractively simple formulation of national policy. It is obscured in that Congress has also provided for class action suits, OFCC's Order No. 4 calls for "relief for members of an 'affected class'," the courts have approved the near-quota of the Philadelphia Plan, and the EEOC has opposed valid employment practices on the grounds that employers did not prove that there was no alternative practice that would also be valid but would provide better racial balance in hiring. Now, from these considerations, it seems that national policy is corrective and requires hiring practices that will maximize the opportunities for employment among groups that have previously been victims of discrimination.

Thorndike (1971) seems to have demonstrated that these two views of national policy--two different definitions of fairness--are inconsistent. The purpose of this preamble is to show that, on the one hand, it calls for maximizing the accuracy of prediction for individuals; and that on the other hand,

it asks for optimizing the relative proportions hired in subgroups. The implications of the research reported here are different for these two interpretations of national policy.

My personal view is that programs of affirmative action (and similar group-referenced policies) are means toward the end of individual equality of opportunity matching equality of qualification. I am here reiterating my earlier view that the basic, or long term, national policy requires that individuals with equal probabilities of success on the job have equal probabilities of being hired (Guion, 1966). I interpret the results of the present study from that frame of reference, and I see three major implications for regulatory agencies.

1. Regulatory agencies should increase their emphasis on job-related constructs. Regulatory policy should be even more concerned than it is with the quality of the thought processes that go into the choice of tests. The competent definition of constructs to be measured, and competence in the choice of valid methods of assessing those constructs, has led to a unique degree of success in this study. In spite of three different kinds of criteria, a multitude of tests, and three different occupations, these investigators have reported significant validities in almost every instance, with some of them being close to magnificent. Contrast this to the usual mixture of some significant and some more nonsignificant validity coefficients, and you reach the conclusion that somebody did something right.

I suggest that part of what was right was an unusual degree of care and intelligence in the selection of tests. It began with job analysis, as both EEOC and OFCC recommend, but it went beyond that. In the first place, the job analysis was done by the investigators themselves so that they could apply their own knowledge of the psychology of human performance to their observations.

Beyond that, the job analysis information was also the basis for criterion development before any final list of tests was approved. These investigators had a rather clear idea of what they wanted to predict before they picked out their predictors.

I have used the term "construct" advisedly. Unlike "idea" or "concept," it refers to a variable that can be rather thoroughly understood. One may first postulate a construct on the basis of a single observation, but a construct grows in precision and meaning as its interrelationships with other variables become more fully known. A well-defined construct is one in which this nomological network is known in some detail. This study's uniqueness derives in part from the use of well-defined constructs, drawn from an analysis of the jobs and applied as predictors. I am not going so far as to suggest that factorial batteries should always be used; I do suggest that tests be chosen on the basis of a great deal of information about what the test has and has not correlated with in the past. Where tests are chosen on this basis, the chances of finding significant empirical validity seem great indeed.

2. The agencies should encourage purification of research. Laboratory research is often criticized as irrelevant, not subject to the vicissitudes of real life, but the laboratory principle of controlling for contaminating error should be observed wherever investigators have the wit and the opportunity to do so. Such control is more evident in these studies than in most validation research.

One example is in the development and use of rating scales. They were carefully constructed to reflect observations of on-the-job behavior. Even more important, the rating process was carefully divorced from administrative procedures. When ratings are used to decide who keeps his job, gets promoted, or wins a raise, their value as research criteria is clouded. Administrative

decisions take into account future expectations, employee needs, organizational needs, and general favoritism--along with performance on the present job. If raters are convinced of the value of the research, and are convinced that their ratings will neither hurt nor help anyone, they can be more honest in their descriptions.

The purification process is further illustrated by the job knowledge and work sample criteria. These represent still more control over contamination, and these criteria are even better predicted. One might complain that the high correlations for predicting job knowledge tests is merely a matter of method variance, but that argument can hardly apply to the two quite different kinds of work samples.

In short, this study suggests that the route to better evidence of validity, and therefore to better knowledge of applicant qualifications, lies in the use of objective, relatively controlled criteria.

3. Regulatory agencies should exercise great caution in demands for evidence of differential validity. In the light of my previously published views (Guion, 1966), the findings of these studies are not personally very satisfying. There is some, but certainly not much, support for a general phenomenon of differential validity. My recommendation from these data is that of an earlier collegiate generation: play it cool. Employers should look for the best evidence that can be found in any given situation, but both they and the agencies should avoid any preconceived ideas of what to expect.

One can find something in these data, if he will ignore other things, to support any preconceived position he likes. If he believes that differential validity is a myth, he can point to the 84 comparisons where there are no significant differences in standard errors, slopes, or intercepts. If he thinks ethnic identification is a moderator, he can point at least to the

seven comparisons where slopes differ significantly and perhaps to the 42 comparisons where intercepts differ. If he thinks testing is disadvantageous to minorities, he can point to some charts where the regression line for the minority group is above that for nonminorities, but if he thinks tests do no harm and may even be biased in favor of minorities, he can point not only to the no-difference comparisons but also to a substantial number where the minority regression line is <u>below</u> that of the whites. However, if he looks at all the data, he sees that patterns from differential validity comparisons are not clear enough for any sort of generalization.

In most of these comparisons, there simply is no evidence of differential validity. Where there is, it appears to work to the disadvantage of the minority group. Moreover, the common pattern (i.e., parallel regression with the minority line lower) has been shown to be possibly a statistical artifact (Linn & Werts, 1971). Differential regressions may be artifactual for other reasons as well; an apparent difference between Mexican-Americans and Caucasians in the Inventory Manager study turned out to be due to differences in the actual tasks performed.

I suspect, therefore, that my recommendation should be stronger. Employers should be required, where technically feasible, of course, to study the possibility of differential validity; it does happen, and in at least one of these comparisons, it was striking. However, a rigorous showing of differential validity should be demanded if the employer expects to act upon the results. In the absence of such a showing, he should pool data for all subgroups so that predictions are based upon the composite sample. Such predictions would be based on more reliable data, and any systematic errors of prediction would probably work to the advantage of members of a disadvantaged group.

I would summarize the information here, and that emerging in the general

literature as well, by suggesting that, as a general rule, the validity of a test against a specified criterion is likely to be about the same for all comers. There are exceptions to the rule, and there are enough exceptions that they must be taken seriously; they are, nevertheless, exceptions.

The rule itself raises an interesting historical question. Testing is not particularly new; employment tests have been validated for over half a century. If a test that was valid for whites was also likely to be valid for blacks, why haven't more blacks been hired? My guess is that minority applicants over the years were routinely rejected regardless of scores, even if tested, because of what was euphemistically called "policy." Test scores were blamed for rejections because that involved less an admission of culpability than does a statement of exclusionary policy; a myth about test unfairness resulted.

To summarize all of this, I believe that the major implication of this study is that procedures used in selection should have at least some validity for at least some people; if they do, and if they are used, then qualified applicants, both minority and nonminority, are likely to be identified. The insistence on some validity for some people will probably do more to usher in an era of genuinely equal opportunity than will the pursuit of the elusive ideal of differential validity.

References

Guion, R. M. Employment testing and discriminatory hiring. Industrial Relations, 1966, 5, 20-37.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

SOURCES OF BIAS IN THE PREDICTION OF JOB PERFORMANCE: IMPLICATIONS FOR FUTURE RESEARCH

S. Rains Wallace

Professor of Psychology

The Ohio State University

Like so many excellent research projects, this study appears to settle some issues while raising others. There are some questions which I would like to hear discussed or to get more information on. I am still worried about the concurrent nature of the study and the degree to which populations have been curtailed on both the predictor and the criterion variables. I take very seriously Dr. Casavantes' concern for the definition of the parent populations themselves. I would like to know how the reliabilities of the criterion measures were determined. I am puzzled by the absence of any analysis of the data provided by the personal history questionnaire, which we are told was administered to all the subjects in the study and would, at a minimum, give us some leads in guessing the degree of range restriction. These are directions for future analysis of the present data, and I assume they will be pursued.

But while all these questions and others are of great interest to me, particularly from the methodological viewpoint, I guess they loom less importantly than might have been the case a few years ago. It appears to me to be about time for us to accept the proposition that written aptitude tests, administered correctly and evaluated against reasonably reliable, unbiased, and relevant criteria, do about the same job in one ethnic group as in another. It seems clear that people like me who expected race to act as a moderator variable for validity relationships were wrong. It seems also clear that people who assumed that all written tests were inappropriate and unfair instruments if applied outside of the WASP culture were equally wrong. In

short, differential or single-group validity is an artifact of small samples, inequalities in restrictions or their correction, or biases in criteria.

It is interesting to note that when we adopt Thorndike's definition of "culture fairness," we are led first to determine the criterion performance of the two populations in question to determine which of them the tests in our battery should discriminate against. Only when there is no difference in performance on the criterion can a "culture free" test be regarded as fair. If, as this study indicates, our more objective and therefore (?) fairer criteria are likely to show poorer black performance, we appear to be stuck with tests upon which black performance is equally low. This is going to cause some unhappiness and much misunderstanding, even as it relieves us of the culture-free absurdity.

However, to the degree that our discussion relates to the use of these validity relationships in the classical selection situation, it may be that this study and others like it are simply too late. In some quarters, at least, the question appears to have shifted from, "Can we use selection tests to select fairly in different ethnic groups?" to "What right has anyone got to select, i.e., reject, at all?"

The difficulty is discussed concisely and dispassionately by Owens and Jewell (1969). They say (pp. 419-420):

> The philosophy that every individual who is capable of work should be placed in a job which demands full and efficient use of his talents can be seen as a rising directional force, exerting increasing pressures on the personnel psychologist to employ methods consistent with this view. Certainly the present selection-rejection model...does not fit comfortably into this philosophical context. The strength of this classic model lies in its provision for probabilistic demonstration that Applicant A is more likely to succeed in a specific job than Applicant B. The model fails, however, to provide information about the skills and abilities of either the selected or rejected applicants as they relate to jobs with different requirements. The selection-rejection model is designed to meet the immediate needs of industry in the most

> efficient (profit-wise) way possible. The needs of the individual and of society are secondary, arguments to the contrary notwithstanding.
>
> There are, however, some very real and immediate manpower problems facing industrial organizations for which the traditional selection-rejection model provides no adequate solutions. The most pressing of these is the shortage of qualified personnel to fill positions at the technical, professional, and managerial levels... At the lower end of the labor market continuum, a different kind of problem exists. For the available unskilled and semiskilled jobs, there is an oversupply of applicants...as the size of this low-level unemployed group grows, both government and industry feel a responsibility to utilize this relatively unused manpower resource ...In addition, there is the humanitarian philosophy founded on the premise that because a person is a human being he deserves the opportunity to realize his talents in activities of his choice--including work.

Thus, we are likely now to hear less about selection, cut-off scores, and the like, and more about diagnosis, finding the right job for a person (whatever the right job is), restructuring jobs so that they make lesser demands on people, or improving our training processes so that anybody can be trained to do anything. This idea has broad appeal in its humaneness and dedication to the total usage of human resources, but one cannot consider it for long before a new question arises, to wit, "How long can an economy survive if the efficiency of its labor force at most strata of difficulty, complexity, and importance is eroded by the placement of workers without regard to their accurately anticipated performance?" If, as these data show, we insist upon placing workers at whatever test performance level in the job of inventory managers, we are going to hire people who, on the basis of any of the criteria examined, perform ineffectively. We cannot escape the fact that this is going to result in much less than optimum management of inventories and that it is going to cost somebody money--namely, you and me, and Charley Brown. It is difficult to be comfortable with the prospect of a society permeated by muddling medical technicians, careless cartographers, misfeasant managers, or pusillanimous policemen. But we are also finding it uncomfortable to refuse

a job to someone who wants it, even if our best prediction is that he will fail out or fail in. We are particularly uncomfortable when this rejection process appears to make things more difficult for one ethnic group than another.

While the way out of this dilemma is not clear to me, a first step in extrication may be to develop more accurate and convincing estimates of the true cost of abandoning the selection process. A second step is the exploration of the possibilities of substantially reducing that cost through other means, e.g., specialized training, advanced supervisory techniques, or extensive job restructure. Obviously, a first requirement for each of these steps is the development of reliable and relevant performance criteria, and it is here that I see this study's major contribution to our thinking about lines for future research.

You can say what you like about supervisors' ratings and I will be glad to help you. The replication in this study of the rater-ratee interaction bias is most convincing, particularly when, on the surface, the supervisors' ratings appeared to be freer of ethnic discrimination than the more objective measures. However, I am constrained to point out, as many of us have for these many years, that the ratings have other faults. Certainly their relevance is open to question (note the low correlation between ratings and work sample in this study), and where they have reliability there is a considerable possibility that it is spurious. Let us hope that this study can provide the cardiac stake and crossroads for the final interment of the supervisory rating criterion so far as research purposes are concerned.

In all fairness, let us also hope for a moratorium on the quest for the philosopher's stone test predictor. Only Pirandello should be expected to enjoy the sight of thousands of tests in search of a criterion. Furthermore, let us ask ourselves a little more carefully what we think we are accomplishing when we validate work samples as predictors against subjective criteria.

Research should be fomented by questions, and I have one which is burning me. Why do the blacks do so poorly on the work sample? There are a number of reasons for thinking this question important. I recognize that many would say that the answer is simple. Their test scores are low, which means their aptitudes are low, ergo their performance is low. Indeed, this is a simple-minded way of saying, with Thorndike, that there is no bias. But somehow this answer fails to content me because if it is true, we seem doomed to reject, with great fairness, many blacks from many jobs. If it is not true, there may be hope of reducing this difference in work performance by recognizing and correcting the factors other than test aptitude that are associated with the apparent inferiority of the blacks (and, indeed, of low test scorers in all ethnic groups).

You have noted and not been surprised by the fact that the written tests predict the job knowledge test criterion better than the other two. There seems only a little difference in the predictive power of the test battery for ratings and the work sample. Remembering that supervisors' ratings correlate more highly with the job knowledge test than the work sample, could we entertain the hypothesis (as Guion did way back in 1965) that the "aptitude" measured by the test battery and associated with the job knowledge test and supervisors' ratings is largely irrelevant to job performance and that, in fact, some set of variables other than aptitude, as we ordinarily define it, is depressing test performance and work sample performance alike? In that case, could we not strive to identify these variables and see what could be done about ameliorating their effects so far as job performance is concerned?

The concurrent nature of this study offers some opportunities along this line. Here I have many questions which I believe could be at least partially answered from the data already at hand. For example, what are the means and variances of time on the job? Of time in employment? Is there a relation

between these variables and performance on the work sample? If not, why not? If there is, what about the relation of experience variables and performance on the predictors? Are there differences among the ethnic groups in terms of experience factors?

If experience is related to work sample performance, shouldn't we expect performance to plateau at some point? If it does, does the ethnic difference remain? Where in the career does the difference appear? Is it constant thereafter?

What about relationships among other variables in the personal history questionnaire and work sample performance. I am substituting for Bill Owens here today, and I would be derelict in my duty if I failed to point out, with others of our speakers, that biographical data may be potent tools not only in improving prediction but also in giving insights into the nature of other measuring instruments such as the work sample. Indeed, there are some things about the population data which, while they give me no insights, certainly give me pause. In the cartographic technician sample which, you will recall, gets most of our attention since all three criterion measures were obtained for it, the black population clearly includes a higher proportion of males, is more experienced, has more "formal education" (whatever that means), and is older. These may be irrelevant facts but, then again, they may not. I would also like to emphasize Dr. Anastasi's mention of the desirability of further exploring the personality traits obtained in the supervisors' ratings.

Finally, I have many questions about the work samples themselves. Some of them are of the more standard quantitative type but most are of a qualitative nature. Can we determine if there are certain aspects of the work sample requirements which account for a major portion of the poorer black performance? Is there reason to believe that the blacks have more difficulty in understanding

the directions? Has any study been made of the relationship between the race of the work sample administrator and the difference in attitude toward the work sample situation? What are the possibilities in the use of work samples in the diagnosis of workers' weaknesses and the provision of remedial training?

I believe that those responsible for this study have made an outstanding contribution by demonstrating that work samples can be constructed and shown to be reliable and facially valid. The very fact that this is possible points to some fascinating lines for future research into the basic nature of work performance. Of course, the question of the relevance and total job coverage of the work sample will be raised. Dr. Anastasi has noted the importance of job analysis and content validity in this connection. It should be possible and desirable also to examine the relationships among the work sample and other objective criteria such as job survival and absenteeism. It would probably be constructive to look at such administratively acceptable but usually unreliable or highly contaminated objective measures as sales, piece-work rate, subordinate performance, etc. The use of critical incident techniques to evaluate the job coverage provided by the work samples seems plausible. We may (however pessimistically) even include administrative evaluations as reflected by promotions and salary levels in our study. But those who reject work samples on the grounds of their lack of credibility or face validity must remember that the selection of criteria is, in the final analysis, always an act of faith. In the light of the evidence for unreliability and bias in supervisors' ratings and the unreliability or contamination in the more real-world type of objective performance measures, the burden of the argument would seem to be on those who attack the work sample rather than on those who defend it. In any case, when one considers what stupid criteria we have been using in our studies of job structure, training effectiveness, supervisory methods, attitudes, motivation,

job satisfaction, compensation, organizational structures, indeed in all of our studies of work in the real world, he is tempted to suggest that we junk it all and start over again with criteria which have sufficient reliability to be themselves susceptible to meaningful study. The construction of large numbers of such criteria and their use in long-term investigations would, I believe, constitute a significant breakthrough in our understanding of what can be accomplished with man--Black, Caucasian, Mexican-American--any man.

References

Guion, R. M. Personnel Testing. New York: McGraw-Hill, 1965.

Owens, W. A., & Jewell, D. O. Personnel Selection. Annual Review of Psychology, 1969, 20, 419-446.